UNIVERSITAT DE BARCELONA
Facultat de Matemàtiques i Informàtica

GRADO DE MATEMÁTICAS

Trabajo fin de grado

CARTOGRAFÍA BAYESIANA DEL CÁNCER EN LA PROVINCIA DE TARRAGONA

Autor: Laura Aixalà Perelló

Director de la UB:	Josep Fortiana
Director del SEPC:	Alberto Ameijide
Realizado en:	Departamento de Matemáticas e Informática

Barcelona, 21 de junio de 2020

A tu Avi.

2010 Mathematics Subject Classification. 62C10, 62F15, 62P10, 65C05, 65C40

Abstract

The goal of this article is to generate cancer risk maps for the region of Tarragona. We use Bayesian hierachical models adapted to the spatial correlation inherent the data. In the report we explain some elements of Bayesian Statistics to set a foundation for the models we apply, we discuss computational techniques and visualize on maps the results we obtained.

Resumen

El objetivo de este trabajo es generar mapas de riesgo de cáncer en la provincia de Tarragona. Empleamos modelos jerárquicos bayesianos adaptados a la correlación espacial que presentan los datos. En la memoria explicamos algunos elementos de estadística bayesiana para fundamentar los modelos aplicados, comentamos las técnicas computacionales y mostramos los resultados obtenidos visualizándolos mediante mapas.

Agradecimientos

En primer lugar, agradecer al Dr. Josep Fortiana por aceptar dirigir este trabajo, por aportar luz en los momentos que parecía que nada tenía sentido y por las horas dedicadas a corregir hasta el último detalle.

También agradecer al *Servei d'Epidemiologia i Prevenció del Càncer de l'Hospital Universitari Sant Joan de Reus* por la oportunidad brindada con la realización de las prácticas que han hecho posible este trabajo, en especial, a Alberto Ameijide por el esfuerzo y el tiempo dedicado durante todas las prácticas, y a Jaume Galceran por hacer posible el convenio.

Y finalmente, quiero dar las gracias a mi familia y a mi pareja por creer en mí durante toda la carrera, y por apoyarme incansablemente a lo largo de todos estos años. Sin vosotros no habría llegado tan lejos. El título será tan mío como vuestro.

Índice

1. Introducción **1**

2. Mapeo de enfermedades **4**

2.1. Estimaciones en áreas pequeñas 4

2.2. Análisis de correlación espacial 5

2.2.1. I de Moran 5

2.2.2. I_i de LISA 7

2.3. Estimaciones por tasas y RIE 10

2.3.1. Tasa bruta 10

2.3.2. Tasas ajustadas por edad 11

2.3.3. Razón de incidencia estandarizada (RIE) 11

3. Estadística Bayesiana **13**

3.1. Inferencia Bayesiana 13

3.1.1. Función de verosimilitud 14

3.1.2. Distribuciónes a priori y a posteriori 15

3.1.3. Distribuciones predictivas 17

3.2. Modelos jerárquicos 17

4. Modelos espaciales bayesianos **18**

4.1. Modelo Besag York Mollie (BYM) 20

4.2. Modelo Suma Ponderada 20

5. Métodos computacionales **23**

5.1. Monte Carlo mediante cadenas de Markov 23

5.1.1. Cadenas de Markov 23

5.1.2. Algoritmos MCMC 24

5.2. `CARBayes` 25

5.2.1. Diagnóstico de convergencia *CARBayes* 25

5.3. OpenBUGS 26

5.3.1. Diagnóstico de convergencia OpenBugs 28

6. Presentación y discusión de resultados **30**

7. Conclusiones y trabajo futuro **38**

8. Apéndice I **43**

8.1. Conceptos básicos de teoría de probabilidad 43

8.1.1. Conceptos previos al Teorema de Bayes 43

8.1.2. Teorema de Bayes 45

8.2. Variables aleatorias 45

1. Introducción

El cáncer es uno de los principales problemas de salud de la población mundial. La Organización Mundial de la Salud (OMS) lo clasifica como la segunda causa de muerte en el mundo. El cáncer fue el responsable de 16.888 muertes en Cataluña en el año 2015 (Clèries et al., 2018), siendo así la segunda causa de muerte en la población catalana después de las enfermedades del aparato circulatorio (Registre de Mortalitat de Catalunya, 2019). En la demarcación provincial de Tarragona se diagnostican más de 4.500 nuevos casos y 1.750 defunciones cada año (Galceran et al., 2013), causando además la pérdida de calidad de vida y sufrimiento de los afectados.

Este trabajo se centra en el análisis de la incidencia de esta enfermedad. La incidencia es el número de nuevos casos de cáncer durante un período especifico de tiempo en una población determinada. Es una medida del grado de exposición que tiene una población a los factores de riesgo de esta enfermedad. La incidencia es la mejor medida para monitorizar tendencias temporales y diferencias geográficas del riesgo de cáncer.

La herramienta fundamental que permite disponer de estos indicadores es el conjunto de registros de cáncer de base poblacional, RCBP. Son sistemas de información que recogen, tratan y almacenan de forma continuada y sistemática datos de todos los casos nuevos de cáncer en una población bien definida (MacLennan et al., 1978). Los RCBPs de diferentes regiones del mundo siguen una amplia lista de principios, criterios y normas de funcionamiento que permiten obtener unos resultados comparables entre ellos. (Jensen et al., 1991; Parkin et al., 1994).

En Cataluña, existen dos Registros de Cáncer de Base Poblacional, el de Tarragona y el de Girona. El Registro de Cáncer de Tarragona (RCT) es un registro gestionado por el Servei d'Epidemiologia i Prevenció del Càncer del Hospital Universitari Sant Joan de Reus. Inició su actividad en el año 1979 y contiene información de base poblacional desde el año 1980.

El RCT sigue los criterios y normas internacionales para la definición de cada caso, y para los sistemas de operación y elaboración de resultados, para poder garantizar la fiabilidad y comparabilidad de los datos con otros registros de cáncer de base poblacional del mundo.

Los datos del Registro de Cáncer de Tarragona han sido publicados en diversas monografías (Galceran et al., 2008; Galceran et al., 2013) y han alimentado las principales publicaciones internacionales sobre incidencia y supervivencia.

Recientemente, el RCT ha publicado un Atlas geográfico sobre la evolución del mapa del Cáncer en los municipios de Tarragona entre los años 1980 y 2009 (Galceran, 2019). Esta publicación sirvió para ver que hay tumores

con patrones geográficos más marcados que podrían ser la clave para crear hipótesis sobre las causas de los tumores en Tarragona. En dicho Atlas se detectó que, en diversos tipos tumorales, el riesgo de desarrollar un cáncer era significativamente superior en los municipios de carácter urbano de la región del Camp de Tarragona. Existen diferentes causas que podrían explicar este hecho: la densidad y tamaño de las poblaciones, la contaminación, la actividad industrial de la zona, el diferente estilo de vida de las zonas urbanas en relación a las zonas rurales, entre otras. Uno de los problemas de dicha publicación fue, que, al trabajar con municipios, el tamaño de Tarragona y Reus, su cercanía de estas dos poblaciones y la heterogeneidad dentro de estas unidades de análisis no permitía acotar las zonas de riesgo para plantear hipótesis sobre las causas de cada tumor, por lo que se planteó un segundo estudio relacionado.

El objetivo de este segundo estudio es rehacer el análisis de riesgo en la provincia de Tarragona, pero por secciones censales, ya que son unidades de análisis más pequeñas y homogéneas. Se realiza mediante un análisis de la distribución espacial de la incidencia de seis de los tumores, en forma de mapas de riesgo. El estudio abarca el período entre los años 2000 y 2014 a lo largo de los diferentes quinquenios.

Este Trabajo de Fin de Grado se centrará en el método usado en este estudio y su base teórica.

El mapeo de enfermedades se usa para explicar y predecir patrones de enfermedades en áreas geográficas, identificar las áreas de mayor riesgo y así proporcionar información para la detección de las causas de dichas enfermedades (Cramb et al., 2016).

El riesgo de padecer una enfermedad se calcula relacionando los casos observados en un área con los casos que se esperaría tener en población estándar equiparable. Estas estimaciones pueden verse afectadas por el tamaño de las poblaciones.

Para suavizar este ruido del riesgo de padecer la enfermedad estudiada, se hace inferencia del riesgo mediante modelos jerárquicos bayesianos (Lawson et al., 2017).

Este tipo de modelos se describen jerárquicamente por etapas y combinan la función de verosimilitud asignada a los datos observados, que es la encargada de introducir esos datos en el modelo, con las distribuciones a priori de los parámetros del modelo. Mediante el Teorema de Bayes se calculan las distribuciones a posteriori de los parámetros con la información de los datos observados (Mesa Páez et al., 2011).

Los modelos jerárquicos bayesianos empleados para calcular de forma robusta el riesgo de cáncer són el de Besag York Mollie y el de Suma Ponderada de Distribuciones a Priori Espaciales de Lawson y Clark (Cramb et al., 2017).

Los métodos bayesianos se han popularizado desde los años ochenta gracias al desarrollo de algoritmos computacionales. Los métodos MCMC son algoritmos que simulan iterativamente la distribución posterior de los parámetros mediante el método de Monte Carlo y las cadenas de Markov (MCMC). .

Recientemente los métodos bayesianos se han aplicado en áreas como la epidemiologia, entre otras, gracias a la aparición de softwares que calculan las simulaciones a posteriori de modelos bayesianos relativamente complejos. El desarrollo de BUGS (Inferencia Bayesiana mediante el algoritmo de muestreo Gibbs) ha sido relevante en la aceptación y generalización del uso de los modelos bayesianos (Lawson, 2018).

Los softwares más comunes són WinBUGS y OpenBugs . En este trabajo se usará el OpenBugs, mediante los paquetes de funciones de R `R2WinBUGS` y `RBrugs`, y el paquete `CARBayes` de funciones de R (Lee, 2013; Galceran, 2019; Vranckx, 2019).

2. Mapeo de enfermedades

El lugar de residencia de un paciente influye en el origen de una enfermedad. La epidemiología espacial cuantifica y explica la variación geográfica de una enfermedad a fin de relacionarla con posibles causas. Con este fin, el mapeo de enfermedades es un componente esencial para esta ciencia ya que permite detectar patrones y áreas de riesgo (Elliot et al., 2004).

Un análisis estadístico espacial consta de tres elementos: los datos observados, el análisis de los datos mediante estadísticos y modelos estadísticos y los mapas de riesgo. Estos tres elementos pueden impulsar el desarrollo posterior de los otros dos ya que cada componente añade información y datos útiles para perfeccionar los demás (Cramb et al., 2016).

Los datos usados para el mapeo de enfermedades tienen un componente geográfico. Este componente puede ser descrito como una ubicación exacta, o, como es más habitual, la región de estudio se divide en diferentes áreas (con formas regulares o irregulares dependiendo del interés o de la información disponible) y se estudia la región en función de estas áreas (Cramb et al., 2016).

En los estudios por áreas se considera que los datos de una misma área tienen una estimación constante. Un método común es el estudio de áreas mediante el conteo de los afectados (Cramb et al., 2016).

2.1. Estimaciones en áreas pequeñas

En este trabajo las áreas de estudio serán las 526 secciones censales de la demarcación provincial de Tarragona. Se medirán las áreas en función de su población y no por su tamaño geográfico. Así, un área grande será un área poblada y un área pequeña una con una población pequeña.

Una población pequeña viene determinada por la enfermedad de interés. Según Cramb (2016), en una enfermedad relativamente poco común como el cáncer, con un porcentaje del 0.5 % de la población afectada, se necesitarían poblaciones de 10.000 habitantes para obtener un número de casos observados funcional para el análisis. Aún así, no hay ninguna definición rigurosa y la selección de una escala espacial apropiada dependerá del objetivo del análisis y de la disponibilidad de los datos.

Los tumores que se han estudiado son los de pulmón, estómago y esófago en hombres y la mama, el esófago y la laringe en mujeres. En la variable *Porcentaje medio* del Cuadro 1 se puede observar el porcentaje medio de población afectada por sección censal de cada tumor. Se han escogido estos seis tumores para el estudio para probar los métodos en diferentes grados de incidencia: siendo alta en el caso de los tumores de pulmón y mama, media

en los tumores masculinos de estómago y esófago y baja en los tumores femeninos de esófago y laringe.

	Min.	1n Qu.	Mediana	Media	3r Qu.	Máx.	Porc. medio
Población M	23.0	453.0	676.0	707.5	932.2	3079.0	-
Población F	10.0	454.0	682.0	696.2	919.0	2496.0	-
Mama F	0.00	6.00	11.00	10.85	15.00	45.00	2.94
Pulmón M	0.00	5.00	8.00	8.46	11.00	37.00	2.09
Estómago M	0.00	0.00	1.00	1.69	2.00	8.00	0.43
Esófago M	0.00	0.00	0.00	0.73	1.00	6.00	0.18
Esófago F	0.00	0.00	0.00	0.11	0.00	2.00	0.03
Laringe F	0.00	0.00	0.00	0.08	0.00	3.00	0.02

Cuadro 1: Resumen de poblaciones e incidencias por sección censal

Las áreas utilizadas en este estudio, con un rango de 10 a 3079 habitantes, son claramente menores que las propuestas por Cramb. Aun así se hizo el estudio con estos datos, ya que eran los que había disponibles para el objetivo del análisis.

2.2. Análisis de correlación espacial

El mapeo de enfermedades está basado en áreas geográficas, por lo que usa datos espaciales que pueden complicar la inferencia estadística debido a la correlación espacial que existe entre áreas cercanas. Esta correlación espacial está explicada por la primera ley de geografía de Tobler dónde se expone que las áreas cercanas son más similares entre si que las que están más separadas (Anselin et al., 2020).

No se pueden hacer análisis de regresión tradicionales ya que la correlación espacial impide suponer que los datos son independientes e idénticamente distribuidos, por lo que se podrían obtener conclusiones falsas (Cramb et al., 2016).

Un método para calcular la correlación global espacial entre áreas de una región es la I de Moran, explicada a continuación. Localmente existe la I_i de LISA (Indicadores Locales de Asociación Espacial), explicada en la sección 2.2.2 más abajo, que es el equivalente local de la I de Moran (Cramb et al., 2016)

2.2.1. I de Moran

La autocorrelación espacial expresa la correlación existente entre las áreas de una región. Su complejidad es debida a que es multidimensional y multidi-

reccional. Se suele medir, de forma global, mediante la I de Moran (Moran, 1950).

La I de Moran se obtiene estandarizando la autocovarianza espacial por la varianza de los datos. Se define como:

$$I = \frac{N}{\sum_i (z_i - \overline{z})^2} \times \frac{\sum_i \sum_j w_{ij}(z_i - \overline{z})(z_j - \overline{z})}{\sum_i \sum_j w_{ij}}, \tag{2.1}$$

donde i, j = 1,...,N y n es el número de áreas de la región. z_i es el valor observado en el área i, y $\overline{z}$ el valor medio observado. w_{ij} es el coeficiente de ponderación (peso) que indica si las áreas i y j son adyacentes/cercanas entre si. En nuestro estudio w_{ij} se define:

$$w_{ij} = \begin{cases} 1, & \text{si las áreas i y j son adyacentes,} \\ 0, & \text{en caso contrario.} \end{cases}$$

El valor esperado de la I de Moran bajo la hipótesis nula de no existencia de autocorrelación espacial es:

$$E(I) = \frac{-1}{N-1}, \tag{2.2}$$

Su varianza es:

$$Var(I) = \frac{NS_4 - S_3 S_5}{(N-1)(N-2)(N-3)(\sum_i \sum_j w_{ij})^2}, \tag{2.3}$$

donde

$$S_1 = \tfrac{1}{2} \sum_i \sum_j (w_{ij} + w_{ji})^2,$$

$$S_2 = \sum_i (\sum_j w_{ij} + \sum_j w_{ji})^2,$$

$$S_3 = \frac{N^{-1} \sum_i (z_i - \overline{z})^4}{(N^{-1} \sum_i (z_i - \overline{z})^2)^2},$$

$$S_4 = (N^2 - 3N + 3)S_1 - NS_2 + 3(\sum_i \sum_j w_{ij})^2,$$

$$S_5 = S_1 - 2NS_1 + 6(\sum_i \sum_j w_{ij})^2,$$

Los valores de I observados que están por debajo de los valores de I esperados indican una autocorrelación espacial negativa. Cuanto mayor sea

la diferencia indicará una dispersión más perfecta de los datos. Los valores de I observados que exceden los valores de I esperados indican una autocorrelación espacial positiva. Cuanto mayor sea la diferencia indicará una correlación más perfecta de los datos. Cuando la diferencia es igual a cero indica un patrón espacial aleatorio, es decir, que no existe correlación espacial entre los valores.

Se han calculado los valores de la I de Moran de los tumores estudiados mediante el paquete de funciones `ape` (Paradis et al., 2020):

```
Moran.I(x, weight, scaled = FALSE, na.rm = FALSE, alternative =
"two.sided")
```

La función `Moran.I` calcula los siguientes elementos bajo la hipótesis nula que no existe correlación:

- observed: La I de Moran calculada.
- expected: El valor de la I de Moran esperado bajo la hipótesis nula.
- sd: la desviación estándar de la I de Moran bajo la hipótesis nula.
- p.value: El P-valor del test de hipótesis nula contra la hipótesis alternativa especificada en `alternative`.

	I observada	I esperada	I sd	I p.value
Mama F	0.0496	-0.0019	0.0265	0.0525
Pulmón M	0.0326	-0.0019	0.0265	0.1936
Estómago M	0.0310	-0.0019	0.0266	0.2153
Esófago M	-0.0115	-0.0019	0.0265	0.7175
Esófago F	0.0412	-0.0019	0.0264	0.1025
Laringe F	-0.0059	-0.0019	0.0259	0.8788

Cuadro 2: I de Moran de cada tumor.

En el Cuadro 2.2 se presentan los resultados de las I de Moran por cada tumor. Los valores observados de todos los tumores tienden a los valores esperados con p-valores mayores a 0.05 por lo que no se rechaza la hipótesis nula de que no existe correlación espacial global entre las secciones censales por ninguno de los tumores.

2.2.2. I_i de LISA

La I_i de LISA, o I_i de Moran Local es un indicador de correlación espacial local, es decir, para cada observación indica como de similares a esta son los valores de las observaciones de su alrededor. La suma de las I_i de LISA de

todas las observaciones es proporcional a un indicador de asociación espacial global (Anselin, 1995).

La I_i de LISA se define como:

$$I_i = \frac{(z_i - \overline{z})}{\sum_{k=1}^{n}(z_k - \overline{z})^2/(n-1)} \times \sum_{j=1}^{n} w_{ij}(z_j - \overline{z})^2, \tag{2.4}$$

donde i, j = 1,...,N y n es el número de áreas de la región. z_i es el valor observado en el área i, y $\overline{z}$ el valor medio observado. w_{ij} es el coeficiente de ponderación (peso) que indica si las áreas i y j son adyacentes/cercanas entre si. w_{ij} también se define:

$$w_{ij} = \begin{cases} 1, & \text{si las áreas i y j son adyacentes,} \\ 0, & \text{en caso contrario.} \end{cases}$$

El valor esperado de la I_i de LISA, bajo la hipótesis nula de no existencia de autocorrelación espacial es:

$$E(I_i) = -\frac{\sum_j w_{ij}}{N-1}, \tag{2.5}$$

Su varianza es:

$$Var(I_i) = \frac{w_{i(2)}(N - S_2)}{N-1} + \frac{2w_{i(kh)}(2S_2 - N)}{(N-1)(N-2)} - \frac{w_i^2}{(N-1)^2}, \tag{2.6}$$

donde

$$w_{i(2)} = \sum_{j \neq i} w_{ij}^2,$$
$$2w_{i(hk)} = \sum_{k \neq i} \sum_{h \neq i} w_{ik} w_{ih},$$
$$S_2 = \frac{\sum_i (z_i - \overline{z})^4}{\sum_i (z_i - \overline{z})^2},$$

El análisis es muy similar al de la I Moran. Los valores de I_i que exceden los valores esperados de I_i indican una autocorrelación espacial positiva entre el valor del área i y los valores de las áreas que están espacialmente a su alrededor. Los valores de I_i por debajo de los esperados de I_i indican una autocorrelación espacial negativa entre el valor del área i y los valores de las áreas que están espacialmente a su alrededor. Cuando la diferencia es igual

a cero indica un patrón espacial aleatorio, es decir, que no existe correlación espacial entre los valores.

Se pueden calcular los valores I_i de LISA de las observaciones de los tumores estudiados mediante el paquete de funciones `spdep` (Bivand et al., 2009):

```
localmoran(x, listw, zero.policy=NULL, na.action=na.fail,
alternative = "greater", p.adjust.method="none", lvar=TRUE,
spChk=NULL, adjust.x=FALSE)
```

	I_i observada	I_i esperada	I_i varianza	I_i sd	p.value
Min.	-5.6216	-0.0019	0.0533	-6.2896	0.0000
1r Qu.	-0.1281	-0.0019	0.1401	-0.3003	0.2902
Mediana	0.0185	-0.0019	0.1969	0.0484	0.4807
Media	0.0326	-0.0019	0.2015	0.1085	0.4669
3r Qu.	0.2301	-0.0019	0.2466	0.5528	0.6180
Máx.	4.8429	-0.0019	0.9921	11.9713	1.0000

Cuadro 3: El resumen de las I_i de LISA del pulmón.

	Min.	1r Qu.	Mediana	Media	3r Qu.	Máx.
Mama F	-2.4612	-0.1245	0.0111	0.0496	0.2261	3.9464
Pulmón M	-5.6216	-0.1281	0.0185	0.0326	0.2301	4.8429
Estómago M	-1.9088	-0.1357	0.0185	0.0310	0.1923	2.6211
Esófago M	-2.5915	-0.2167	-0.0036	-0.0115	0.1791	4.5452
Esófago F	-1.9667	-0.0953	0.1127	0.0412	0.1127	4.7761
Laringe F	-2.6015	-0.0806	0.0677	-0.0059	0.0677	6.3846

Cuadro 4: I_i valores observados de LISA de cada tumor.

Como la variable N (N=526) es constante por todos los tumores tendremos que el valor esperado no varia en función del tumor estudiado ($E(I_i)$ = -0.0019).

En el Cuadro 3 se observa que el rango de los valores observados de I_i del pulmón va desde -5.6216 hasta 4.8429. Este rango de valores indica que hay secciones censales i, cuyos valores observados están correlacionados positivamente, otras negativamente, y también otras que no tienen correlación espacial en relación a los valores de sus secciones censales vecinas.

En el Cuadro 4 se pueden observar comportamientos parecidos en los demás tumores.

2.3. Estimaciones por tasas y RIE

El mapeo de enfermedades estudia las características de una población heterogénea respecto a aspectos sociodemográficos, como, por ejemplo, la edad y el género de sus habitantes. Las poblaciones se suelen estudiar como un conjunto de subgrupos. Por consiguiente, cualquier medida o estadístico general refleja el valor de esa medida en cada uno de los subgrupos en los que se ha dividido la población. Ese valor se define mediante la media de los valores para los grupos individuales, ponderados por sus tamaños relativos. En consecuencia, el tamaño del subgrupo será proporcional a su influencia en la media (Schoenbach, 1999).

Los estadísticos más sencillos son aquellos que usan de forma directa los datos observados en la población. Los datos observados en el estudio de riesgo de cáncer es un listado del número de casos incidentes de cada sección censal.

La limitación del conteo de casos reside en la dificultad al comparar las áreas entre si al ser la población heterogénea y por tanto diferentemente distribuida en cada área. Por esa razón se utilizan variables que reflejan el riesgo relativo de padecer una enfermedad por cada sección censal y no el recuento de casos en sí mismo. El riesgo relativo se calcula a partir tasas (Galceran, 2019).

2.3.1. Tasa bruta

Una tasa bruta es aquella que no tiene en cuenta de forma explícita la composición de la población. Es la forma más sencilla y directa de resumir una característica de una población.

La tasa bruta utilizada en este estudio es la proporción entre el número de casos en un determinado período de tiempo y el tamaño de la población de riesgo en ese mismo período. Por ello se supone que el riesgo permanece constante en todas las categorías de edad y sexo.

La tasa bruta del área i, se calcular como:

$$TB_i = \frac{O_i}{P_i} \times k, \tag{2.7}$$

donde O_i son los casos observados en en el área i y P_i el número de personas residentes en el área i.

La tasa bruta de una región generalmente se expresa como el número de casos de la región por cada año y cada 100.000 personas, es decir, $k = 100.000$.

Las tasas brutas en enfermedades que varían en función de la edad pueden cambiar si se calculan fijando la variable edad.

2.3.2. Tasas ajustadas por edad

La tasa ajustada por edad ajusta la estructura de edad de la población a partir de una población estándar de referencia (p.e. mundial, europea).

La tasa ajustada por edad es una medida resumen que considera que la población estudiada tiene la misma estructura de edad que la población estándar. Se suelen agrupar los casos en m grupos de edad, generalmente 18 grupos de cinco años consecutivos (0-4, 5-9,. . . , 85+). También suele expresarse por cada 100.000 personas, es decir, $k = 100.000$.

$$TA = \frac{\sum_{i=1}^{m} w_i TEE_i}{\sum_{i=1}^{m} w_i} \times \frac{O}{O - SE} \times k, \tag{2.8}$$

donde TEE_i son las tasas especificas por edad (tasas brutas de un grupo de edad específico), w_i son los coeficientes de ponderación de cada grupo de edad en la población estándar, O es el número de casos observados en el período estudiado y SE el número de casos sin edad conocida en el mismo período.

Este método directo permite la comparación de áreas, pero requiere recuentos específicos por cada área que podrían no estar disponibles o ser inestables. Otro inconveniente de este método es la definición de la población estándar, que puede ser significativa o no con nuestra población de estudio y estimar erróneamente la tasa.

2.3.3. Razón de incidencia estandarizada (RIE)

El estandarizado indirecto se basa en estimar el número de casos que se esperaría si la población siguiera las mismas tasas de contagio que la población estándar. Este procedimiento busca facilitar la comparación de las medidas de resumen entre áreas. El cociente calculado mediante este proceso es la Razón de incidencia estandarizada (RIE).

La razón de incidencia estandarizada (RIE) es un estimador de riesgo relativo de un área i. Este calcula el riesgo de un área en relación a una población estándar.

$$RIE_i = \frac{O_i}{E_i}, \tag{2.9}$$

donde O_i es el número de casos observados en el área i, y E_i son los casos esperados en el área i, donde:

$$E_i = \sum_{jk} P_{ijk}\lambda_{jk}, \tag{2.10}$$

donde P_{ijk} es la población del área i, del grupo de edad j en el año k, y λ_{jk} es la tasa de incidencia de la población de referencia del grupo de edad j y año k.

En el presente estudio las áreas son las secciones censales de la Provincia de Tarragona y la población de referencia es la total de la provincia.

$$\sum_i P_{ijk},$$

Las tasas de incidencia se calcularan para cada combinación de tipos tumoral y sexo en la población de referencia del grupo de edad j y año k. Así tenemos que:

$$\lambda_{ijk} = \frac{\sum_i O_{ijk}}{\sum_i P_{ijk}}, \tag{2.11}$$

donde O_{ijk} es el número de casos incidentes del área i, en el grupo de edad j y año k.

El problema principal de este método es que en áreas muy pequeñas las estimaciones brutas de la RIE son muy inestables e imprecisas debido a que su varianza es inversamente proporcional a los valores esperados. Por otra parte, la variación de los casos observados suele ser mayor a los esperado, produciendo lo que se llama extra-variabilidad.

Las estimaciones con menos casos tienden a tener RIEs extremas en el mapa. En caso que esto ocurra, se utilizan modelos de suavizado para obtener estimaciones de la RIE de cada área más robustas. Estos modelos "suavizan" los valores de la RIE de cada área en función de sus áreas vecinas. Los modelos que se emplean son modelos jerárquicos bayesianos, que vemos a continuación.

3. Estadística Bayesiana

La metodología bayesiana está basada en la interpretación subjetiva de la probabilidad, es decir, puede contener la opinión y grado de convicción del observador. Esta subjetividad se mide mediante rangos de variación y la simulación de diferentes condiciones de incertidumbre.

En Estadística Bayesiana toda cantidad desconocida (como los parámetros de un modelo) se trata como una variable aleatoria. Los parámetros se definen mediante distribuciones a priori, definidas también con parámetros aleatorios, estableciendo una jerarquía de parámetros (Lawson, 2018). El considerar los parámetros variables aleatorias permite introducir estructuras de correlación específicas, en nuestro caso espaciales (Lee et al., 2018). Esta estructura incorpora la información contenida en los datos mediante el Teorema de Bayes (Clark et al., 2006).

Teorema 3.1 (Fórmula de Bayes)**.** *Sean* y, $\theta \subset \mathcal{A}$ $p(\theta)$, $p(y)>0$. *Entonces:*

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}. \tag{3.1}$$

El objetivo de la inferencia bayesiana es combinar las distribuciones previas de los parámetros, que incluye la interpretación subjetiva del observador, con un modelo estadístico, que contiene la propia observación de los datos. La combinación de las distribuciones previas de los parámetros y el modelo estadístico permite obtener nuevas distribuciones a posteriori de los parámetros con dicha interpretación del observador y información de los datos, es decir, más informativas (Mesa Páez et al., 2011).

Los modelos jerárquicos bayesianos son aquellos que se describen en diferentes etapas mediante las cuales se va añadiendo la información que se tiene de los datos al modelo con distribuciones a priori. Suelen dividirse en tres etapas: en la primera se describen los datos observados y, mediante el modelo estadístico, en la segunda se describen los parámetros de los efectos aleatorios del modelo, θ y en la tercera los hiperparámetros del modelo, λ (Cramb et al., 2017). La distribución a posteriori de los parámetros θ se calcula mediante el teorema de Bayes:

$$p(\theta|y,\lambda) = \frac{p(y|\theta)p(\theta|\lambda)}{p(y|\lambda)}. \tag{3.2}$$

3.1. Inferencia Bayesiana

La inferencia estadística es el conjunto de métodos que permiten sacar conclusiones de una población o muestra. La inferencia bayesiana es una infe-

rencia estadística basada en la distribución de probabilidad a posteriori que resulta una vez se han observado las dadas.

La inferencia bayesiana trata el parámetro θ como una variable aleatória y lo infiere a partir de los datos y mediante (Cramb et al., 2016):

1. Un modelo $p(y|\theta)$; función de verosimilitud.
2. Una distribución a priori de θ, $p(\theta)$.

La combinación de la función de verosimilitud con la distribución a priori mediante el Teorema de Bayes da una distribución a posteriori del parámetro a partir de la cual se basa la inferencia:

$$\text{d. a posterior} \propto \text{d. a priori} \times \text{f. de verosimilitud.}$$

3.1.1. Función de verosimilitud

La función de verosimilitud es la función de densidad de probabilidad de las variables observables condicionada a un valor dado de los parámetros (Lawson, 2018).

La función de verosimilitud de los datos $\{y_i\}$, i = 1,...,m, se define como:

$$L(y|\theta) = \prod_{i=1}^{n} f(y_i|\theta), \tag{3.3}$$

también es útil la versión logarítmica del modelo:

$$l(y|\theta) = \sum_{i=1}^{n} log f(y_i|\theta), \tag{3.4}$$

donde θ es un vector de parámetros y f$(\cdot|\cdot)$ es la función de densidad o la función de masa.

Notaremos la función de verosimilitud cómo $p(y|\theta)$.

Principio 3.2 (Principio de verosimilitud)**.** Toda la información de los datos que es relevante para las inferencias sobre el valor de los parámetros del modelo se encuentra en la función de verosimilitud (estrictamente hablando, en su clase de equivalencia, salvo una constante multiplicativa).

El Principio de verosimilitud implica que toda la información existente de los datos se encuentra en la función de verosimilitud y, además, también implica que si existe alguna información que no aparece en el principio de

verosimilitud entonces esta no afecta a la inferencia ya que sino existiría información no contenida en la función de verosimilitud (Lawson, 2018).

En la función de verosimilitud se asume que los valores de y dados por los parámetros son condicionalmente independientes, dados los valores de los parámetros. Entonces, la relación entre observaciones procede del hecho de compartir parámetros en un nivel superior de la jerarquía (Lawson, 2018).

3.1.2. Distribuciónes a priori y a posteriori

La distribución a priori de un parámetro θ es la función de probabilidad o la función de densidad de probabilidad que expresa la probabilidad de cada valor de θ antes de observar una muestra y, es decir, refleja el conocimiento previo de θ antes de observar los datos. Otra interpretación de la distribución a priori es que sirve para añadir conocimiento adicional al modelo y así obtener una distribución del parámetro más realista (Lawson, 2018).

La distribución a priori de la variable θ se nota cómo $p(\theta)$ y se escoge de forma subjetiva en base a estudios previos y/o conocimiento de los expertos.

Cuando la integral de $p(\theta)$ sobre su rango Ω no es finita, se dice que la distribucion a priori es impropia:

$$\int_\Omega p(\theta)d\theta = \infty.$$

Las distribuciones a posteriori deben ser propias, aunque provengan de distribuciones a priori impropias. Si una distribución prior da lugar a una posterior impropia, no sirve para el análisis y debe ser reemplazada (Lawson, 2018).

Las distribuciones a priori informativas son las distribuciones que añaden información sobre el parámetro. La elección de una distribución a priori se suele hacer en función de lo que se sepa del comportamiento y el rango del parámetro. Si se carece de información previa se usa una distribución a priori no informativa.

Las distribuciones a priori no informativas se eligen de forma que interfieran lo mínimo posible con los datos. Tienden a ser relativamente planas. Por ejemplo, si se busca una distribución a priori con valores positivos se suelen usar distribuciones de la familia gamma, gamma inversa o uniforme. Por esa razón, en uno de los modelos se ha usado la distribución $Gamma(1, 0.01)$ que tiene una media de 1 y una varianza bastante grande de 100. Para las desviaciones estándar se suelen usar distribuciones uniformes de rango grande. Los parámetros de regresión de rango infinito acostumbran a usar distribuciones centradas en el cero con una gran varianza, una distribución normal puede ser una buena candidata (Lawson, 2018).

La función de verosimilitud y las distribuciones a priori proporcionan dos

tipos de información diferente en cualquier análisis. La función de verosimilitud informa sobre el parámetro a través de los datos, mientras que las distribuciones a priori informan a través de creencias o suposiciones previas al análisis. Cuando el tamaño de la muestra es grande, la función de verosimilitud contribuye más al cálculo del riesgo relativo, mientras que cuando la muestra es pequeña, las distribuciones a priori dominan el análisis (Lawson, 2018).

La distribución a posteriori de un parámetro θ se nota como $p(\theta|y)$ y refleja todo el conocimiento que se tiene de θ una vez hechas las asunciones previas y observados los datos. Por ese motivo, la distribución a posteriori es el objeto que se usa para hacer la inferencia.

La distribución a posteriori de θ se calcula mediante el Teorema de Bayes:

$$p(\theta|y) = \frac{p(y,\theta)}{C} = \frac{p(y|\theta)p(\theta)}{C}, \tag{3.5}$$

dónde

$$C = \int_{\Omega} p(y|\theta)p(\theta)d\theta = p(y). \tag{3.6}$$

En consecuencia, la distribución a posteriori se puede expresar cómo:

$$p(\theta|y) \propto p(y|\theta)p(\theta).$$

Ciertas combinaciones de funciones de distribuciones a priori con funciones de verosimilitud conducen a distribuciones a posteriori de la misma familia que la distribución a priori.

Una distribución a priori es conjugada cuando la distribución a priori y la posteriori del parámetro son de la misma familia. Es decir, los parámetros de la distribución a posteriori están en función de los parámetros de la distribución a priori y los datos (Lawson, 2018).

F. verosimilitud	D. a priori	D. a posteriori
$y \backsim Poisson(\theta)$	$\theta \backsim Gamma(\alpha, \beta)$	$\theta\|y \backsim Gamma(\sum y_i + \alpha, m + \beta)$
$y \backsim Binomial(p, 1)$	$p \backsim Beta(\alpha_1, \alpha_2)$	$p\|y \backsim Beta(\sum y_i + \alpha_1, m - \sum y_i + \alpha_2)$
$y \backsim Normal(\mu, \tau)$, $\tau\ fijada$	$\mu \backsim Normal(\alpha_0, \tau_0)$	$\mu\|y \backsim Normal(\frac{\tau_0 \sum y_i + \alpha_0 \tau}{m\tau_0 \tau}, \frac{\tau_0 \tau}{m\tau_0 + \tau})$
$y \backsim Gamma(1, \beta)$	$\beta \backsim Gamma(\alpha_0, \beta_0)$	$\beta\|y \backsim Gamma(1 + \alpha_0, \beta_0 + \sum y_i)$

Cuadro 5: Ejemplos de resultados de conjugaciones.

La conjugación puede no ser posible en modelos con una jerarquía de parámetros compleja. En el caso de los modelos jerárquicos usados en el mapeo de enfermedades es poco probable encontrar una conjugación.

3.1.3. Distribuciones predictivas

La distribución posterior resume el conocimiento de los datos observados. Sin embargo, también existen otras distribuciones útiles para hacer predicciones de datos no observados, actuales o futuros. La distribución predictiva posterior de $\overline{y}$ es el resultado de "marginalizar" es decir, multiplicar por la densidad de probabilidad posterior de los parámetros, $p(\theta|y)$, e integrar, con respecto a estos parámetros, la distribución de verosimilitud de los nuevos datos, $p(\overline{y}|\theta)$, para definir la contribución de los datos observados a la predicción:

$$p(\overline{y}|y) = \int p(\overline{y}|\theta)p(\theta|y)d\theta. \tag{3.7}$$

3.2. Modelos jerárquicos

Los modelos jerárquicos permiten especificar modelos estocásticos mediante etapas. Estas etapas se construyen a partir de relaciones locales simples, pero que en total permite explicar el comportamiento general. Esta definición del modelo por etapas conduce a una estructura jerárquica dónde cada etapa define la relación entre los datos observados y los parámetros desconocidos (Cramb et al., 2017).

Los modelos jerárquicos se pueden resumir en tres etapas:

- Etapa 1: Función de verosimilitud de los datos dependiendo de unos parámetros.
- Etapa 2: Los parámetros tienen cada uno su función de distribución a priori.
- Etapa 3: La distribución a priori pueden depender de ulteriores parámetros que se pueden llamar distribuciones hiperprioris.

El enfoque bayesiano permite desarrollar modelos complejos y realistas a partir de relaciones condicionales simples. Este enfoque permite relajar la condición de independencia en los datos a una independencia condicional. De esta forma se puede añadir la dependencia de la correlación espacial entre las áreas, generalmente en la segunda o la tercera etapa del modelo, y suele funcionan adecuadamente en estudios de áreas pequeñas con baja incidencia (Lawson, 2018).

4. Modelos espaciales bayesianos

Los mapas de riesgo de los datos observados sin tratar o suavizar suelen tener ruido que dificulta su interpretación (Besag, 1986). Para identificar patrones uniformes de los datos basados en covariables y sus factores espaciales subyacentes, se utilizan modelos bayesianos jerárquicos (Gerber et al., 2015).

Los mapas de riesgo suavizados se construyen a partir de las distribuciones predictivas posteriores de los datos.

El objetivo de este estudio es hacer mapas de riesgo dónde el riesgo se mide con la variable RIE:

$$RIE_i = \frac{O_i}{E_i}. \tag{4.1}$$

Para obtener los valores de la RIE suavizados, los casos observados del área i, O_i, pasarán a ser el valor medio de la distribución predictiva a posteriori de los datos observados, y los casos esperados del área i, E_i, se mantendrán igual.

Al "pintar" los mapas se tuvo que escoger un valor concreto de la distribución predictiva posterior de los datos, y se escogió la media. Al perderse mucha información al reducir toda la información dada por la distribución de los valores a su media, se calculó y "pintó" la probabilidad de que el riesgo fuera mayor que 1, es decir que hubiera más casos observados que esperados, por cada sección censal y tumor. Esta variable permite medir la fiabilidad de los valores de la RIE y fue nombrada PRP.

Los modelos Bayesianos usados para mapear la incidencia del cáncer en áreas pequeñas siguen una estructura jerárquica en tres etapas (Cramb et al., 2017):

Etapa 1:	$y_i \backsim Poisson(E_i e^{\mu_i})$
Etapa 2:	$\mu_i = \alpha + x_i^T \beta + R_i$
Etapa 3:	$\alpha \backsim p(\cdot \mid \lambda_\alpha)$
	$\beta \backsim p(\cdot \mid \lambda_\beta)$
	$R_i \backsim p(\cdot \mid \lambda_R)$

Cuadro 6: Modelo jerárquico de tres etapas

La etapa 1 es el modelo de probabilidad. En esa etapa se asigna una distribución de probabilidad a los datos observados. La distribución de Poisson es apropiada cuando el número de casos observados es bajo comparado con la población. A veces, la distribución binomial se prefiere para áreas pequeñas debido a que la distribución de Poisson tiene cierta probabilidad de obtener

más casos que personas en cada área. Sin embargo, este hecho es extremadamente improbable en enfermedades raras donde la diferencia práctica entre las distribuciones de Poisson y binomial es insignificante. Otras veces, se prefiere usar la distribución binomial negativa ya que la distribución de Poisson restringe la media para que sea igual a la varianza (Cramb et al., 2017).

En los modelos empleados se ha asignado una distribución de Poisson al número de casos incidentes $\{y_1, \ldots, y_N\}$ ya que se está estudiando una enfermedad relativamente poco común.

La etapa 2 es la expresión del riesgo relativo logarítmico de μ_i. El parámetro μ_i se suele expresar en forma de ecuación de regresión dónde α es el efecto aleatorio fijo general, β son los coeficientes de regresión de las covariables y x_i denota el vector de las covariables relacionadas con los datos por cada área i. Por último, R_i representa los efectos aleatorios del modelo con estructura espacial. Estos efectos aleatorios espaciales se pueden formar a partir de múltiples componentes (Besag et al. 1991).

En la etapa 3 se definen las distribuciones a priori para cada uno de los parámetros del riesgo relativo. Se puede añadir una cuarta etapa a la jerarquía con las distribuciones hiperpriori de los parámetros λ_α, λ_β o λ_R .

En los modelos que se han usado para el estudio, se les ha asignado una distribución uniforme al efecto aleatorio fijo α y una distribución normal a los efectos de las covariables de media 0 y distribuciones hiperpriori gammas inversas para las desviaciones estándar.

Para aplicar un suavizado espacial al modelo, se puede asignar una distribución autoregresiva condicional, CAR, o alternativa, a los efectos aleatorios R_i. Estas distribuciones representan la autocorrelación espacial en datos relacionados con un conjunto de unidades de área no superpuestas (Lee, 2013).

Existen suavizados espaciales globales y locales. Los globales, contrariamente a los locales, aplican consistentemente los mismos parámetros de suavizado en toda la región. Los modelos globales CAR son fáciles de implementar, pero al no diferenciar entre áreas pueden sobresuavizar áreas adyacentes discontinuas. Los locales tienen en cuenta estas posibles discontinuidades entre áreas (Cramb et al., 2017).

El modelo jerárquico con suavizado global usado para estudiar la incidencia del cáncer es el Besag York Mollie, y el modelo con suavizado local es el modelo Suma Ponderada de Distribuciones Espaciales a Priori de Lawson y Clark. Las distribuciones de los efectos espaciales de ambos modelos se describen a continuación.

4.1. Modelo Besag York Mollie (BYM)

El modelo BYM es intrínsecamente CAR, ICAR (Cramb et al., 2017). Este modelo asigna las siguientes distribuciones condicionales a los parámetros aleatorios con estructura espacial de los modelos espaciales jerárquicos anteriormente descritos:

$$R_i = S_i,$$

$$S_i|s_{\setminus i} \backsim N(\frac{1}{\sum\limits_j w_{ij}} \sum_j w_{ij} s_j, \frac{\sigma_s^2}{\sum\limits_j w_{ij}}),$$

dónde w_{ij} es el elemento de la fila i y columna j de la matriz W de las ponderaciones espaciales que determina la proximidad entre los efectos aleatorios. Los elementos de W se suelen definir de forma binaria:

$$w_{ij} = \begin{cases} 1, & \text{si las areas i y j son adyacentes,} \\ 0, & \text{en caso contrario.} \end{cases}$$

Este modelo implica que los efectos aleatorios espaciales estructurados S_i son la media de los efectos aleatorios de las áreas vecinas.

En caso que existan efectos aleatorios espaciales no estructurados U_i, tendríamos que $R_i = S_i + U_i$ dónde $U_i \backsim Normal(0, \sigma_U^2)$. Este modelo se conoce como el modelo de convolución.

A las varianzas comunes σ_U^2 y σ_s^2 se les asignaran distribuciones gamma inversa.

En consecuencia, tenemos que en los modelos basados en CAR, la fuerza de la autocorrelación parcial depende del número de áreas adyacentes.

4.2. Modelo Suma Ponderada

El modelo de Lawson y Clark es una extensión del modelo BYM de convolución que permite detectar discontinuidades (Galceran, 2019). En este modelo los efectos aleatorios espaciales se definen:

$$R_i = p_i S_i + (1 - p_i) Z_i + U_i.$$

El parámetro Z modela las discontinuidades abruptas entre áreas. La distribución a priori de Z establecida por Lawson y Clark se basa en que existe un riesgo absolutamente diferente entre áreas vecinas, y se define como:

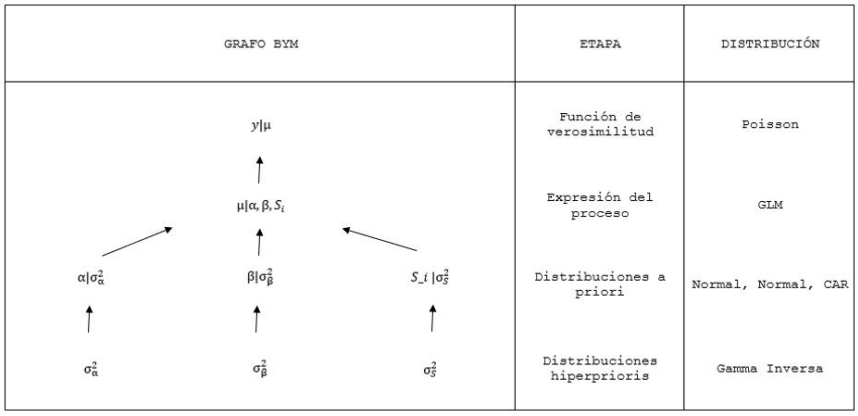

Figura 1: Modelo Besag, York, Mollie. Las variables y su dependencia se muestra en el grafo BYM. También se muestran las etapas del modelo jerárquico y sus respectivas distribuciones.

$$\pi(Z_1, ..., Z_n) \propto \frac{1}{\sqrt{\lambda}} exp(-\frac{1}{\lambda} \sum_{i \backsim j} |Z_i - Z_j|),$$

dónde λ actúa como constante restrictiva y se distribuye como una gamma inversa.

Observamos que si $p_i = 1 \ \forall i$ se obtiene el modelo BYM que no admite discontinuidades, mientras que si $p_i = 0$ el modelo es totalmente discontinuo.

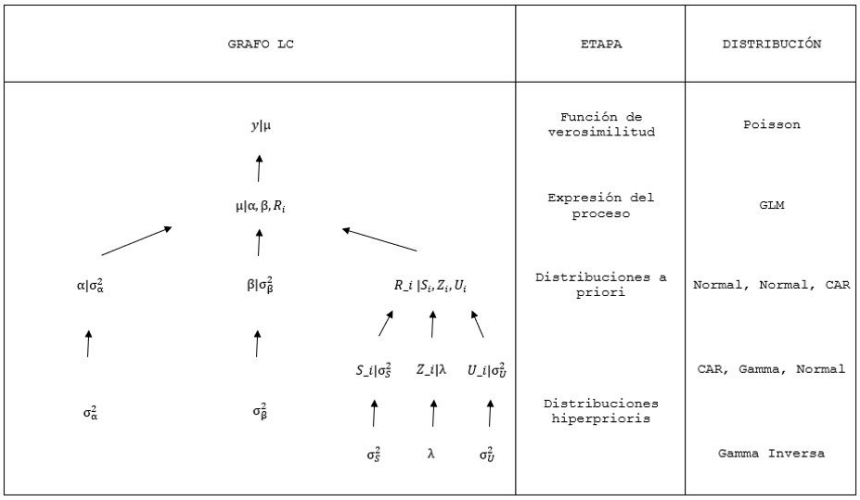

Figura 2: Modelo Suma Ponderada de Lawson y Clark. Las variables y su dependencia se muestra en el grafo LC. También se muestran las etapas del modelo jerárquico y sus respectivas distribuciones.

5. Métodos computacionales

Para que los modelos bayesianos jerárquicos usados en el mapeo de enfermedades sean realistas deben tener más de dos etapas y, consecuentemente, distribuciones a posteriori de los parámetros complejas. Para simular las distribuciones a posteriori de estos parámetros se necesitan algoritmos de muestreo (Lawson, 2018). Los métodos MCMC (Método de Monte Carlo mediante cadenas de Markov) son algoritmos que, dada una distribución de probabilidad, simulan iterativamente la distribución a posteriori de los parámetros mediante cadenas de Markov.

5.1. Monte Carlo mediante cadenas de Markov

La inferencia bayesiana se basa en el comportamiento de la distribución a posteriori. Para obtenerla, se necesita la distribución marginal de los datos que se define mediante una integral, que no siempre es posible de calcular. En estos casos, se utilizan métodos computacionales que aproximan la distribución a posteriori. El método de Monte Carlo estima integrales definidas finito-dimensionales.

El método MCMC consiste en generar una cadena de Markov cuya distribución límite sea la distribución a posteriori que se quiere modelar.

En los siguientes apartados se definirán los principales conceptos para poder entender ese proceso.

5.1.1. Cadenas de Markov

Para definir una cadena de Markov se tienen que definir previamente una serie de conceptos. Estos conceptos se han extraído del trabajo de final de grado de Cristina Rosich (2018).

Definición 5.1. ***Proceso estocástico***
Sea $\mathcal{T}$ un subconjunto de $[0, \infty)$. Un proceso estocástico es una familia de variables aleatorias $\Theta = \{\Theta\}_{t \in \mathcal{T}}$. Si $\mathcal{T} = \mathbb{N}_0$ el proceso estocástico es discreto, si $\mathcal{T} = [0, \infty)$ el proceso estocástico es continuo.

Una cadena de Markov es un proceso estocástico con la propiedad de Markov (ecuación 5.1). Esta propiedad sustenta que los estados futuros solo dependen del estado presente y no de los pasados.

Definición 5.2. ***Cadena de Markov***
Sea E un conjunto de estados. Sea $\Theta = \{\Theta\}_{t \in \mathbb{N}_0} \in E$ un proceso estocástico. $\Theta = \{\Theta\}_{t \in \mathbb{N}_0} \in E$ es una cadena de Markov con

- *Distribución inicial* $\lambda := (\lambda_\theta, \theta \in E)$,
  donde $\lambda_\theta = P(\Theta_0 = \theta)$.
- *Matriz de transición* $M := (p_{ij}, \theta_i, \theta_j \in E)$,
  donde $p_{ij} = P(\Theta_{n+1} = \theta_j | \Theta_n = \theta_i)$.

Si cumple que para todo $n \in \mathbb{N}_0$ *y todo subconjunto de estados* $\{\theta_0, ..., \theta_{n+1}\} \subset E$, *si* $P(\Theta_n = \theta_n, ..., \Theta_0 = \theta_0) > 0$, *entonces*

$$P(\Theta_{n+1} = \theta_{n+1} | \Theta_n = \theta_n, ..., \Theta_0 = \theta_0) = P(\Theta_{n+1} = \theta_{n+1} | \Theta_n = \theta_n) = p_{n,n+1}. \tag{5.1}$$

Definición 5.3. ***Cadena de Markov irreducible***
Una cadena de Markov Θ *es irreducible si desde cualquier estado* $\theta_i \in E$ *se puede acceder a cualquier otro estado* $\theta_j \in E$.

Definición 5.4. ***Proceso estocástico estacionario***
Un proceso estocástico Θ *es estacionario si por todo subconjunto de tiempo* $\{t_1, ..., t_n\} \subset \mathcal{T}$ *y todo subconjunto de estados* $\{\theta_0, ..., \theta_{n+1}\} \subset E$, *entonces* $\forall h \in \mathcal{T}$*:*

$$P(\Theta_{t_1+h} = \theta_1, ..., \Theta_{t_n+h} = \theta_n) = P(\Theta_{t_1} = \theta_1, ..., \Theta_{t_n} = \theta_n). \tag{5.2}$$

En particular, una cadena de Markov Θ *es estacionaria si, y solo si, la distribución marginal de* Θ_n, $P(\Theta_n = \theta)$, *no depende de n.*

Bajo ciertas condiciones, una cadena de Markov converge a su distribución estacionaria:

$$lim_{t \to \infty} P(\Theta_t = \theta | \Theta_0 = \theta_0) = p(\theta). \tag{5.3}$$

Definición 5.5. ***Distribución de equilibrio***
Sea Θ *una cadena de Markov, su distribución inicial* λ *es invariante si* $\lambda M = \lambda$. *En este caso se dice que la cadena de Markov tiene una distribución de equilibrio o distribución límite.*

5.1.2. Algoritmos MCMC

Los métodos MCMC son métodos de simulación que generan muestras de las distribuciones a posteriori y las cantidades de interés a posteriori. Estos métodos simulan valores sucesivamente a partir de una densidad propuesta, no necesariamente similar a la densidad a posteriori.

Para simular la distribución a posteriori $p(\theta|y)$ se simula una cadena de Markov $\theta_1, \theta_2, ...$ cuya distribución estacionaria es $p(\theta|y)$. Es necesario simular la cadena un número elevado de iteraciones para aproximarse a la distribución estacionaria. Se eliminan los primeros valores "de calentamiento" de la simulación porque no están en el estado estacionario.

Cada valor simulado de la cadena, θ_t, depende solo de su predecesor, θ_{t-1}, por la propiedad de Markov.

Si el algoritmo se implementa correctamente, la cadena será convergente independientemente de los valores iniciales.

Los métodos MCMC se implementan mediante softwares que calculan las simulaciones a posteriori de modelos bayesianos. Existe una amplia gama de softwares para implementar modelos con distribuciones de riesgo espaciales, y métodos basados en modelos jerárquicos. Los softwares que se han utilizado permiten implementar modelos intrinsicamente CAR y pueden usarse dentro del software R. Se han utilizado OpenBUGS, mediante los paquetes de funciones de *R* `R2WinBUGS` y `BRugs`, y el paquete de funciones de *R* `CARBayes`.

5.2. `CARBayes`

Un lenguaje de programación diseñado para simulaciones de modelos jerárquicos bayesianos es BUGS, implementado en los programas JAGS, WinBUGS y OpenBUGS, accesibles desde el entorno R de tratamiento de datos. Estos programas están explicados en el siguiente siguiente apartado. En el caso del modelo BYM se ha utilizado el paquete de funciones de *R* `CARBayes` por su facilidad de uso en comparación con el software BUGS. El paquete de funciones de *R* `CARBayes` permite especificar fácilmente la información de contigüidad espacial como una matriz binaria. También, dada una matriz de vecindad, el modelo BYM se puede implementar mediante una sola llamada de la función `S.CARbym` en *R* y, realiza las simulaciones considerablemente más rápido que el software BUGS (Lee, 2013).

5.2.1. Diagnóstico de convergencia *CARBayes*

Para analizar las simulaciones obtenidas mediante el método MCMC se utilizan gráficos de diagnóstico. Estos gráficos son representaciones visuales de las simulaciones MCMC. En el análisis se han utilizado tres funciones diferentes del paquete de funciones `bayesplot` de *R* (Gabry, 2020) que crean diferentes gráficos de la simulación de los parámetros del modelo. Se ha utilizado el tumor de pulmón en hombres como ejemplo.

Primero se han representado los trace plots (gráficos de traza) mediante la función `mcmc_trace`. Estos gráficos indican la regularidad y la cobertura del soporte de la distribución a posteriori por parte de la muestra.

Los gráficos de traza de la simulación de la incidencia de casos de cáncer

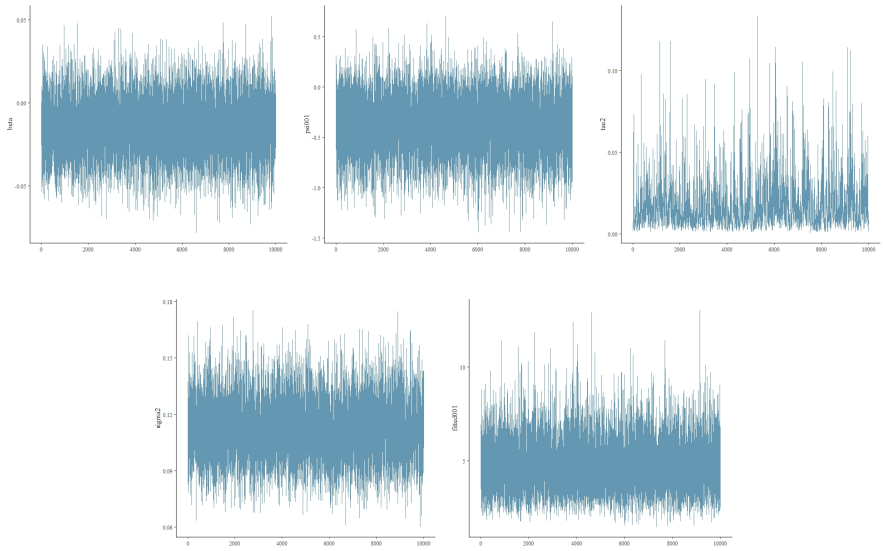

Figura 3: Gráficos de la función `bayesplot::mcmc_trace` para la simulación del pulmón. Las variables de los gráficos son por orden: beta, psi001, tau2, sigma2, fitted001.

de pulmón presentan una regularidad aceptable. Básicamente parecen ruido aleatorio saltando alrededor de un número relativamente constante. Cuando no se comportan bien el trazado no es uniforme (p.e. puede parecer una onda sinusoidal ruidosa, puede saltar entre dos estados ruidosos o simplemente pasar de uno a otro por un corto período de tiempo y luego saltar otra vez).

En los segundos gráficos se representa la autocorrelación de los parámetros mediante la función `mcmc_afc`. El gráfico de autocorrelación para cada parámetro ilustra el grado de correlación entre las muestras de MCMC separadas por diferentes rezagos. Se busca que la autocorrelación descienda rápidamente a cero, comportamiento que puede apreciarse en los gráficos de la Figura 4.

Los gráficos de la función `mcmc_dens` en la Figura 5 representan la densidad de las distribuciones posteriores marginales de los parámetros.

5.3. OpenBUGS

OpenBUGS es un software para el análisis bayesiano de modelos estadísticos complejos que ejecuta las simulaciones de los métodos de Monte Carlo

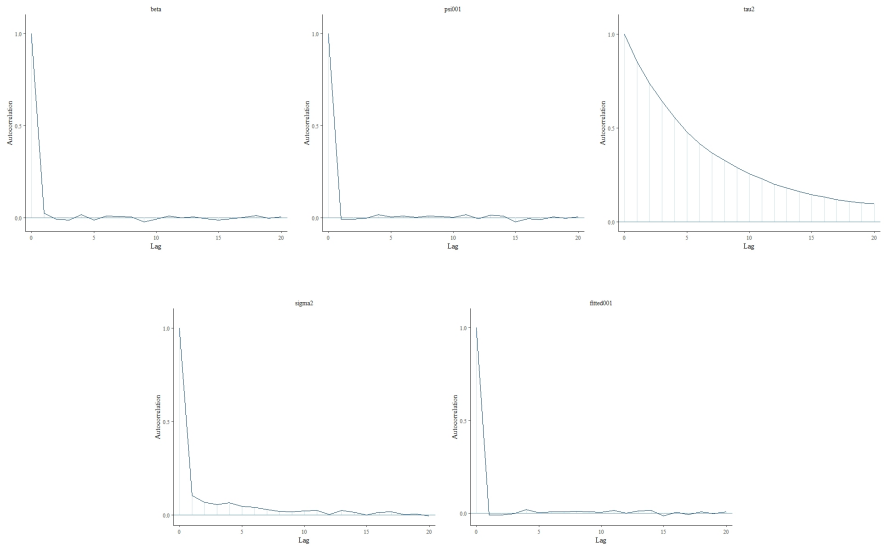

Figura 4: Gráficos de la función `bayesplot::mcmc_acf`. Las variables de los gráficos son por orden: beta, psi001, tau2, sigma2, fitted001.

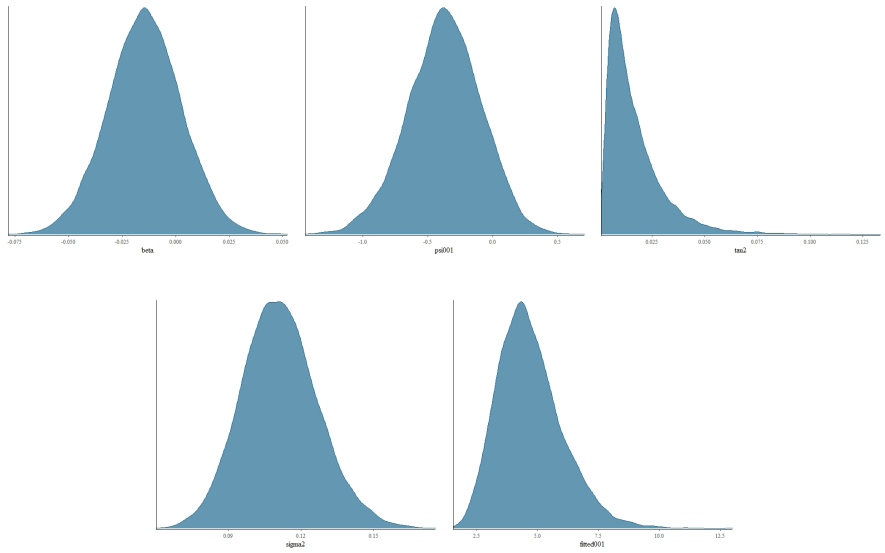

Figura 5: Gráficos de la función `bayesplot::mcmc_dens`. Las variables de los gráficos son por orden: beta, psi001, tau2, sigma2, fitted001.

con cadenas de Markov (MCMC). Implementa el muestreo Gibbs (Gibbs sampling). Como JAGS, puede ejecutarse en múltiples sistemas operativos, mientras que el WinBUGS solo puede usarse en Windows. Se puede usar como una aplicación independiente o bien a través de R mediante los paquetes de funciones `R2WinBUGS` y `BRugs`. Hemos usado OpenBUGS (en vez

de JAGS) para aprovechar porciones de código preexistente en la institución que soporta el trabajo.

OpenBUGS implementa las simulaciones MCMC y las "muestra" de acuerdo con los criterios definidos. Se han generado tres cadenas de 70.000 iteraciones, con un calentamiento de 10.000 iteraciones y un intervalo de 60. El modelo de Lawson y Clark se ha implementado mediante la función `bugs` epecificando el modelo en la variable `model.file`.

5.3.1. Diagnóstico de convergencia OpenBugs

Los gráficos de diagnóstico utilizados para analizar las simulaciones obtenidas mediante el software OpenBugs están explicados en el apartado *Diagnóstico de convergencia CARBayes*.

Los trace plots (gráficos de traza) producidos por la función `mcmc_trace` de la muestra, obtenida mediante el software OpenBugs, de la simulación del modelo de Lawson y Clark por el tumor de pulmón, presentan una regularidad aceptable:

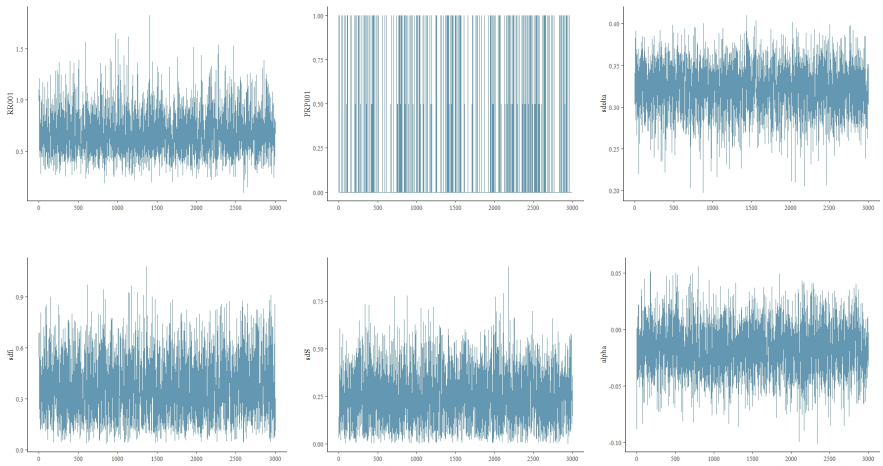

Figura 6: Gráficos de la función `bayesplot::mcmc_trace` para la simulación del pulmón. Las variables de los gráficos son por orden: RR001, PRP001, sdelta, sfi, sdS, alpha.

Los gráficos, producidos por la función `mcmc_acf`, que representan la correlación entre las muestras, decrecen rápidamente a cero indicando que las muestras MCMC obtenidas son independientes.

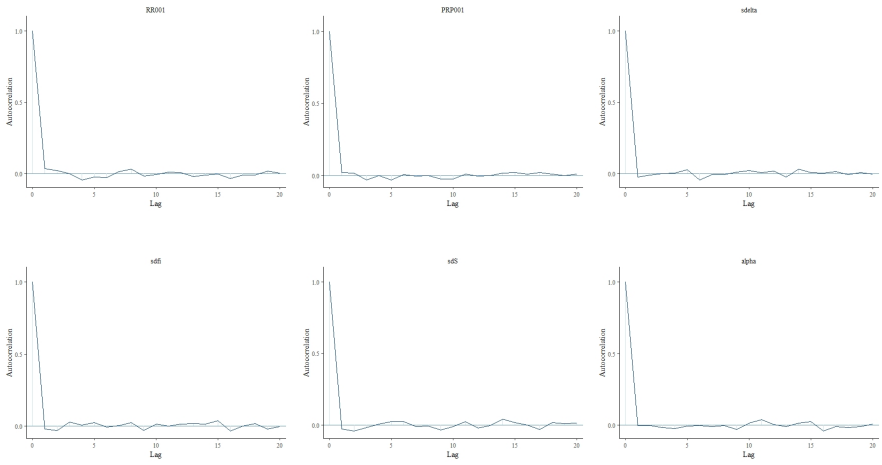

Figura 7: Gráficos de la función `bayesplot::mcmc_acf` para la simulación del pulmón. Las variables de los gráficos son por orden: RR001, PRP001, sdelta, sfi, sdS, alpha.

Y a continuación, se muestran los gráficos de las densidades de las distribuciones posteriores marginales de los parámetros del modelo de Lawson y Clark, producidos por la función `mcmc_dens`:

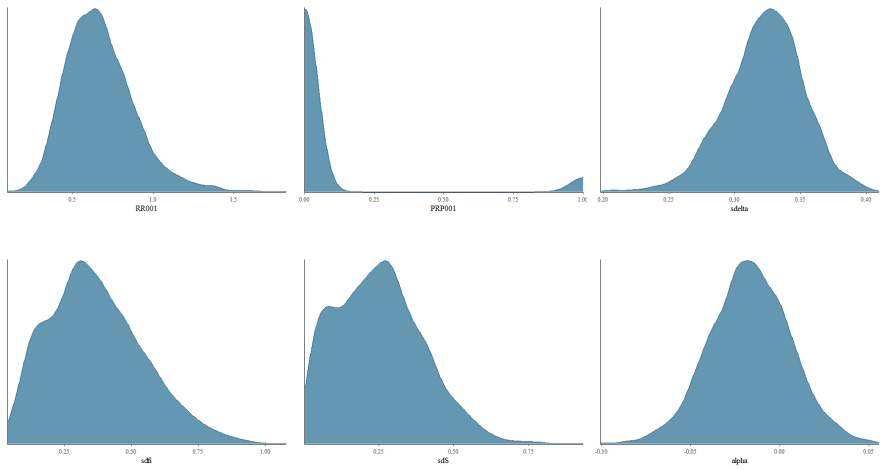

Figura 8: Gráficos de la función `bayesplot::mcmc_dens` para la simulación del pulmón. Las variables de los gráficos son por orden: RR001, PRP001, sdelta, sfi, sdS, alpha.

6. Presentación y discusión de resultados

El principal objetivo de este proyecto es crear los mapas de riesgo de incidencia de cáncer, para seis tumores, de las secciones censales de la provincia de Tarragona en el período de tiempo que abarca desde el año 2000 hasta el año 2014.

Mediante los softwares `CARBayes` y `OpenBUGS` se han calculado los valores de la RIE y los valores de la probabilidad de que la RIE sea mayor que 1, es decir, de la variable PRP, que mide la fiabilidad de los valores de la RIE obtenidos.

Los valores de la variable RIE y la variable PRP se han clasificado en diferentes rangos a los cuales se les han asignado colores para poder visualizarlos en un mapa. Los mapas se han creado mediante la función `spplot`.

Los puntos de corte de los rangos de colores de los mapas de la variable RIE son 0 / 0,50 / 0,666 / 0,75 / 0,833 / 0,90 / 0,95 / 0,99 / 1,01 / 1,05 / 1,10 / 1,20 / 1,333 / 1,5 / 2 / 2,5. Los colores más cercanos al verde oscuro indican un menor riesgo de incidencia de cáncer en la sección censal, mientras que los colores más cercanos al granate un mayor riesgo de incidencia de cáncer.

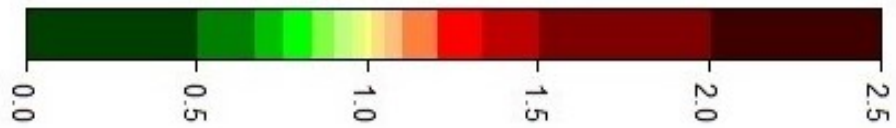

Figura 9: Rangos de los colores de la RIE

```
spplot(Mapa,"SIR",at=c(0, 0.50, 0.666, 0.75, 0.833, 0.90, 0.95,
     0.99, 1.01, 1.05, 1.10, 1.20, 1.333, 1.5, 2, 2.5), col.
    regions=c("#003F00", "#007F00", "#00BF00", "#00FF00", "#7
    FFF3F", "#BFFF7F", "#DFFF7F", "#FFFF7F", "#FFDF7F", "#
    FFBF7F", "#FF7F3F", "#FF0000", "#BF0000", "#7F0000", "#3
    F0000"))

# Mapa es la variable que contiene los valores de la RIE y PRP
    de cada seccion censal.
```

Los puntos de corte de los rangos de colores de los mapas de la variable PRP son 0 / 0,10 / 0,20 / 0,40 / 0,60 / 0,80 / 0,90 / 1,001. Los colores más cercanos al verde indican menor probabilidad, mientras que los colores más cercanos al rojo mayor probabilidad.

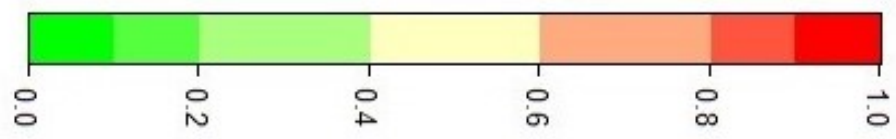

Figura 10: Rangos de los colores de la PRP

```
spplot(Mapa,"PRP",at=c(0, 0.10, 0.20, 0.40, 0.60, 0.80, 0.90,
    1.001), col.regions=colorRampPalette(c("green", "#FFFFBF" ,
    "red"))(7))
```

Los mapas obtenidos se muestran a continuación. Han sido agrupados en figuras de cuatro mapas. Los dos mapas superiores son los obtenidos con el modelo BYM, a la izquierda el que representa los valores de la variable RIE y a la derecha los de la variable PRP. Los dos mapas inferiores son los obtenidos con el modelo Suma Ponderada de Lawson y Clark, con el mismo orden de variables. En cada figura se describe el tipo de tumor representado.

Se puede observar como en los tumores con una incidencia más elevada, es decir, los cánceres de mama y pulmón, el modelo utilizado, ya sea el BYM o el de Lawson y Clark, no presenta diferencias significativas en los mapas, indicando que los valores obtenidos de la variable RIE son parecidos.

En el resto de tumores se puede observar que los mapas de la variable RIE alcanzan valores más extremos mediante el modelo de Lawson y Clark que con el modelo BYM, con menor notoriedad en los tumores de "media" incidencia y mayor en los de "baja" incidencia.

En los mapas de la variable PRP de "media" incidencia, es decir, los cánceres de estómago y esófago en hombres, los valores del modelo BYM y el modelo de Lawson y Clark se asemejan en algunas de las secciones censales con valores más extremos, pero, en general, los valores de la variable PRP calculados con el modelo de Lawson y Clark tienden a ser más cercanos a 0.5, es decir, con el modelo de Lawson y Clark se obtienen resultados menos informativos.

En los mapas de la variable PRP de "baja" incidencia, es decir, los cánceres de laringe y esófago en mujeres, los valores de la variable PRP calculados con el modelo de Lawson y Clark tienden a ser menores de 0.5, resultado que no se corresponde con los valores de la variable PRP calculados con el modelo BYM.

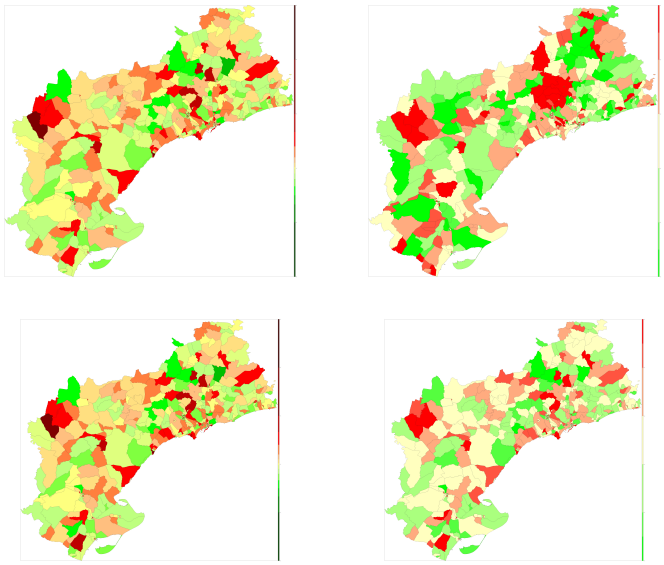

Figura 11: Mama en mujeres. Arriba BYM CARBayes (RIE y PRP). Abajo Lawson y Clark OpenBugs (RIE y PRP).

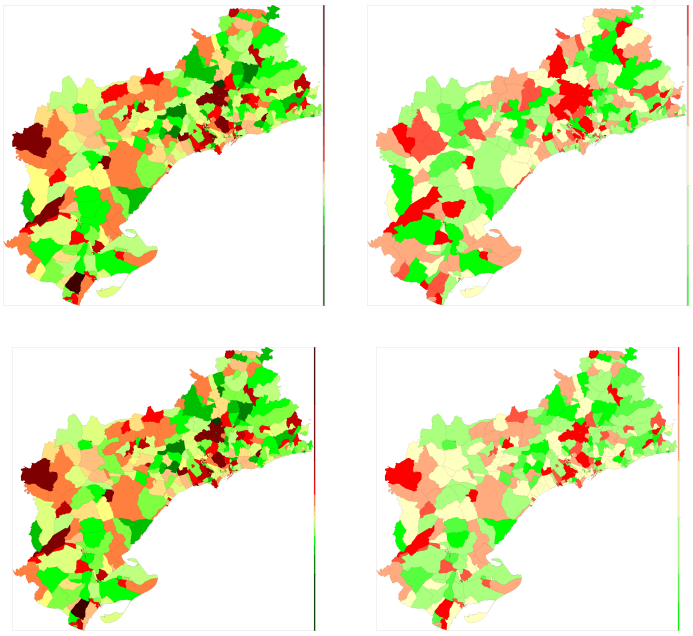

Figura 12: Pulmón en hombres. Arriba BYM CARBayes (RIE y PRP). Abajo Lawson y Clark OpenBugs (RIE y PRP).

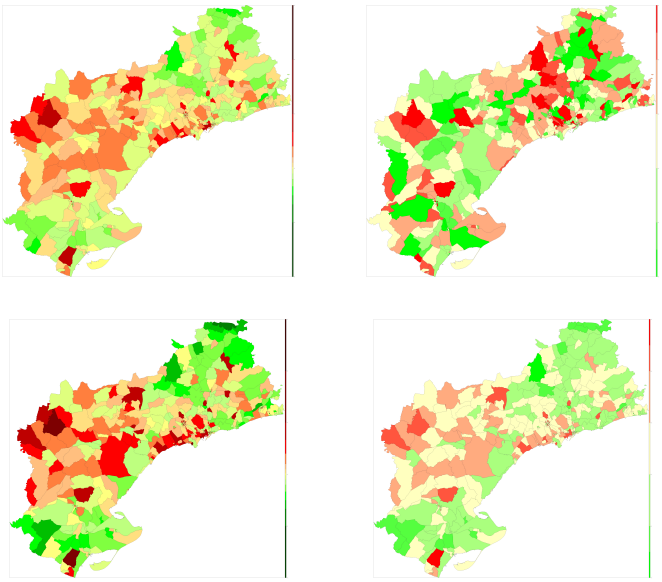

Figura 13: Estómago en hombres. Arriba BYM CARBayes (RIE y PRP). Abajo Lawson y Clark OpenBugs (RIE y PRP).

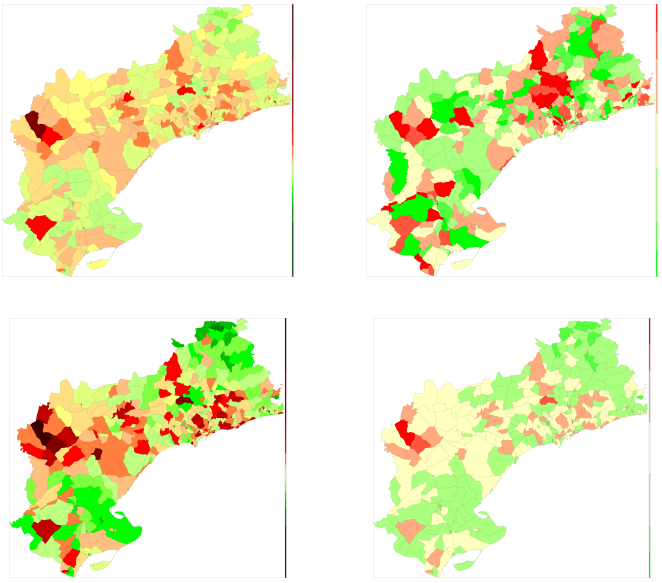

Figura 14: Esófago en hombres. Arriba BYM CARBayes (RIE y PRP). Abajo Lawson y Clark OpenBugs (RIE y PRP).

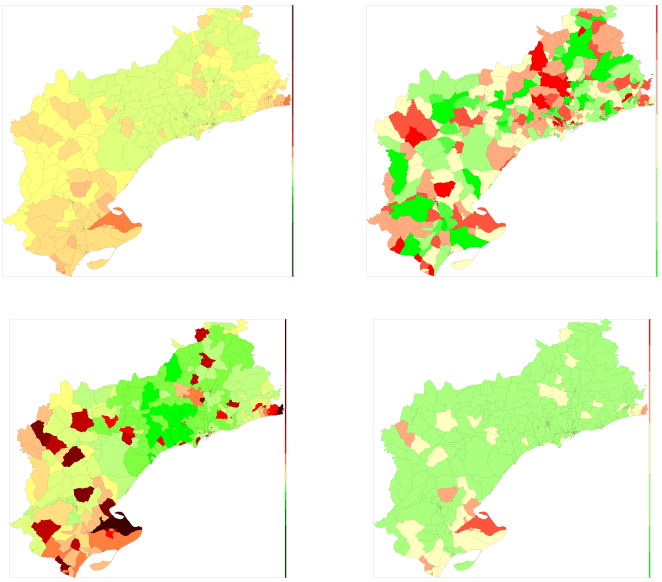

Figura 15: Esófago en mujeres. Arriba BYM CARBayes (RIE y PRP). Abajo Lawson y Clark OpenBugs (RIE y PRP).

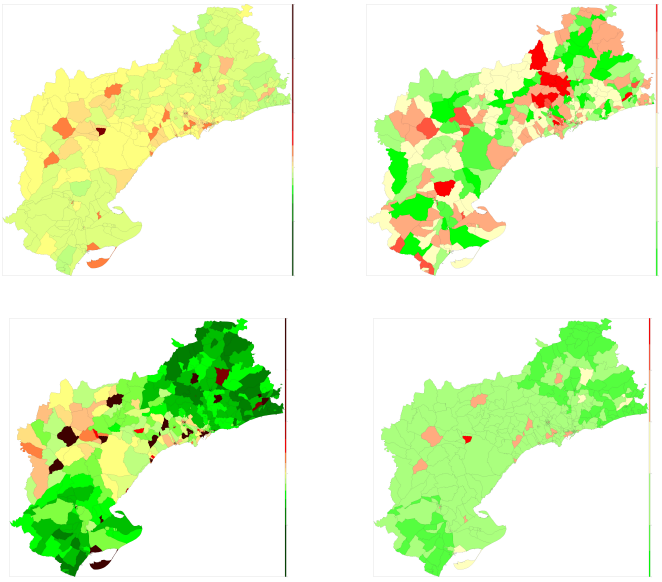

Figura 16: Laringe en mujeres. Arriba BYM CARBayes (RIE y PRP). Abajo Lawson y Clark OpenBugs (RIE y PRP).

Como se ha explicado en la introducción, en la publicación sobre la evolución del mapa del Cáncer en los municipios de Tarragona entre los años 1980

y 2009 (Galceran, 2019) se detectó que el riesgo de desarrollar cáncer era significativamente superior en los municipios de carácter urbano de la región del Camp de Tarragona.

Las Figuras 17, 19, 20, 21 i 23 contienen los mapas de riesgo de cáncer en municipios de dicha publicación de los seis tumores estudiados por secciones censales. El tumor de pulmón es el tumor que más evidencia este hecho de los seis que se escogieron para probar el método en secciones censales.

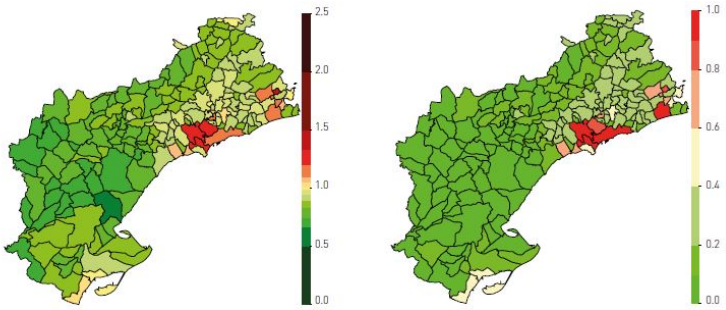

Figura 17: Pulmón en hombres por municipios entre el período 2000 - 2009 (Galceran et al., 2019). Mediante el modelo de Lawson y Clark OpenBugs (RIE y PRP).

La Figura 18 contiene las ampliaciones de los mapas por secciones censales en el área del Camp de Tarragona del tumor de pulmón. Estas ampliaciones muestran variaciones dentro de los municipios de Reus y Tarragona, así como del Camp de Tarragona en general, permitiendo acotar las áreas de riesgo para plantear hipótesis más acertadas sobre las posibles causas de cada tumor.

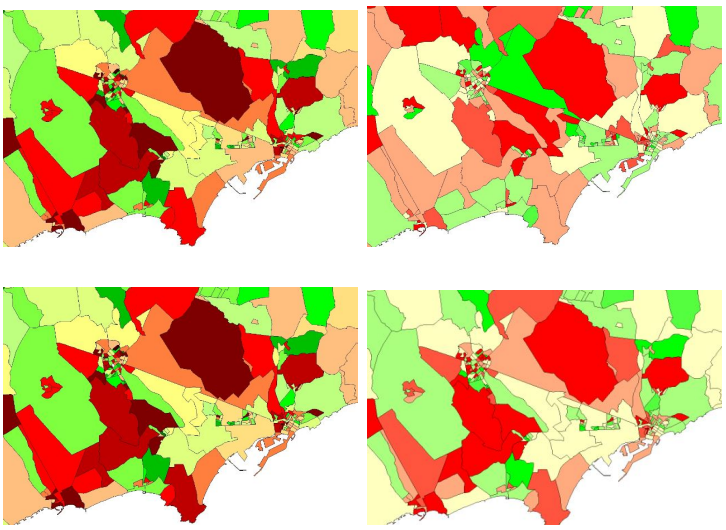

Figura 18: Pulmón en hombres en el Camp de Tarragona. Arriba BYM CAR-Bayes (RIE y PRP). Abajo Lawson y Clark OpenBugs (RIE y PRP.)

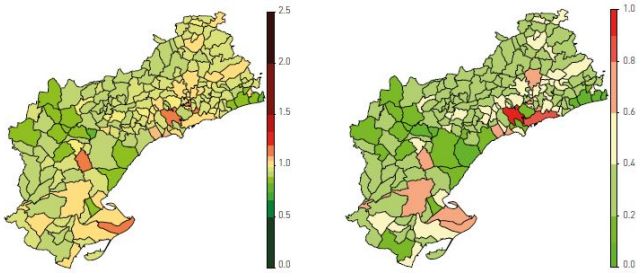

Figura 19: Mama en mujeres por municipios entre el período 2000 - 2009 (Galceran et al., 2019). Mediante el modelo de Lawson y Clark OpenBugs (RIE y PRP).

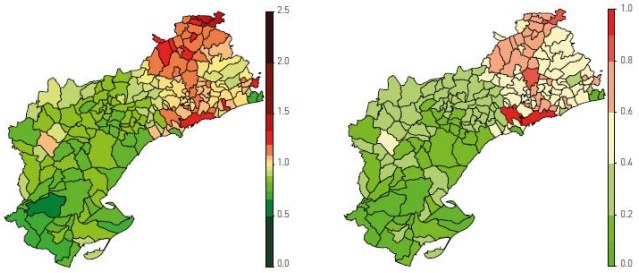

Figura 20: Estómago en hombres por municipios entre el período 2000 - 2009 (Galceran et al., 2019). Mediante el modelo de Lawson y Clark OpenBugs (RIE y PRP).

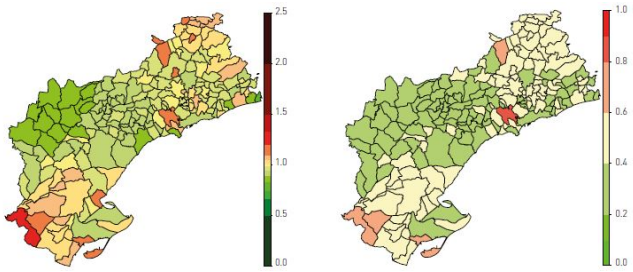

Figura 21: Esófago en hombres por municipios entre el período 2000 - 2009 (Galceran et al., 2019). Mediante el modelo de Lawson y Clark OpenBugs (RIE y PRP).

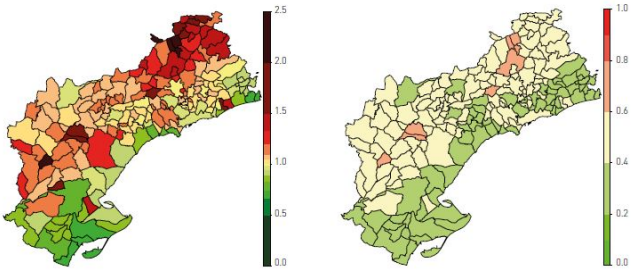

Figura 22: Esófago en mujeres por municipios entre el período 2000 - 2009 (Galceran et al., 2019). Mediante el modelo de Lawson y Clark OpenBugs (RIE y PRP).

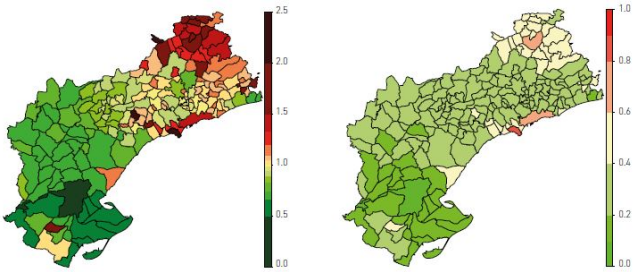

Figura 23: Laringe en mujeres por municipios entre el período 2000 - 2009 (Galceran et al., 2019). Mediante el modelo de Lawson y Clark OpenBugs (RIE y PRP).

7. Conclusiones y trabajo futuro

El objetivo de este trabajo es generar mapas de riesgo de cáncer en la provincia de Tarragona. El riesgo de cada sección censal se ha calculado mediante la razón de incidencia estandarizada (RIE) suavizada mediante modelos jerárquicos bayesianos.

Los datos observados se introducen al modelo mediante la función de verosimilitud, que describe la dependencia de los parámetros, de los valores de la muestra y contiene toda la información de los datos relevante para la inferencia.

Los datos que tienen componentes geográficas presentan correlación espacial entre áreas cercanas. Esta característica se incluye en el modelo mediante las distribuciones a priori, con estructuras espaciales, de los parámetros.

La inferencia de los datos se hace a partir de la distribución a posteriori de los parámetros. A partir de las distribuciones a posteriori de los parámetros se calcula la distribución predictiva posterior de los datos que resume el conocimiento de los datos observados. Los valores suavizados de la RIE se calculan mediante las medias de la distribuciones predictivas posteriores de los datos.

Los modelos jerárquicos bayesianos que se han empleado son el de Besag, York, Mollié (BYM) y el modelo Suma Ponderada de Lawson y Clark. El modelo de Lawson y Clark es una extensión del modelo BYM que permite discontinuidades.

Las distribuciones posteriores de los parámetros de los modelos se han simulado a través de los métodos de Monte Carlo mediante Cadenas de Markov (MCMC), por medio del software OpenBUGS a través de los paquetes de funciones de *R* `R2WinBUGS` y `BRugs` en el modelo de Lawson y Clark, y el paquete de funciones `CARBayes` de *R* en el modelo BYM.

Los mapas de la variable RIE obtenidos a partir de los valores calculados con los softwares se han acompañado de los mapas de la variable PRP que mide la fiabilidad de los valores de la RIE obtenidos.

El mapeo de riesgo de cáncer por secciones censales ha permitido detectar las áreas con más riesgo. Ambos modelos han generado mapas muy parecidos en tipos tumorales con incidencia elevada. Sin embargo, en tipos tumorales de media y baja incidencia los mapas presentan valores más extremos mediante el modelo Suma Ponderada de Lawson y Clark. No obstante, con ambos métodos se pueden obtener conclusiones parecidas.

En lo puramente computacional, una continuación inmediata de este proyecto es la implementación de los modelos en JAGS y en Stan para una mejor eficiencia y, para datos más voluminosos, en Tensorflow-Probability.

En lo metodológico, ampliando los modelos para tener en cuenta también la dimensión temporal de los datos y otros factores como, por ejemplo, el nivel socio-económico de los habitantes de cada sección censal.

Este estudio no está cerrado y quedan cuestiones por resolver como, por ejemplo, el porque los mapas obtenidos por municipios y los mapas obtenidos por secciones censales son aparentemente opuestos en algunos de los tumores.

Para finalizar, recalcar la importancia de los registros de enfermedades que posibilitan este tipo de estudios. Este trabajo se enmarca en un servicio epidemiológico provincial. De tener registros regionales o nacionales, este trabajo podría servir como piloto para estudios de mayor alcance, así como también de otras enfermedades.

Referencias

[1] Anselin, L., Li, X. (2020). Tobler's Law in a Multivariate World. Geographical Analysis, April, 1–17. https://doi.org/10.1111/gean.12237.

[2] Anselin, L. (1995). Local Indicators of Spatial Association—LISA. Geographical Analysis, 27(2), 93–115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x.

[3] Besag, J. (1986). On the Statistical Analysis of Dirty Pictures.

[4] Bivand, R., Altman, A., Assunção, R., Berke, O., Bernat, A., Blanchet, G., Blankmeyer, E., et al. (2009). Package "spdep." In Computational Statistics and Data Analysis. https://doi.org/10.1016/j.csda.2008.07.021

[5] Clark, J. S., Gelfand, A. E. (2006). Hierarchical Modelling for the Environmental Sciences: Statistical Methods and Aplications. https://doi.org/10.1017/CBO9781107415324.004.

[6] Clèries R., Ameijide A., Marcos-Gragera R., Pareja L., Carulla M., Vilardell M.L., Esteban L., Buxó M., Espinàs J.A., Puigdefàbregas A., Ribes J., Izquierdo A., Galceran J., Borrás J.M. (2018). Predicting the cancer burden in Catalonia between 2015 and 2025: the challenge of cancer management in the elderly. Clin Transl Oncol. 2018 May;20(5):647-657.

[7] Cramb, S., Duncan, E., Baade, P., Mengersen, K. (2017). Investigation of Bayesian spatial models. 9, 1–109.

[8] Cramb, S., Duncan, E., White, N., Baade, P., Mengersen, K. (2016). Spatial Modelling Methods.

[9] Elliott, P., Wartenberg, D. (2004). Spatial epidemiology: Current approaches and future challenges. 112(9), 998–1006. https://doi.org/10.1289/ehp.6735.

[10] Gabry, J., Mahr, T., Bürkner, P.-C., Modrák, M., Barrett, M. (2020). Package ' bayesplot .' 1. https://doi.org/10.1111/rssa.12378>.The

[11] Galceran J., Ameijide A., Carulla M., Bigorra J., Saladié F. (2019). L'evolució del mapa del càncer a Tarragona, 1980-2009 Registre de Càncer de Tarragona, Fundació per a la Investigació i Prevenció del Càncer.

[12] Galceran J., Gumà J., Carulla M., Ameijide A., Saladié F., Borràs J. (2013). El càncer a Tarragona, 2013. Dades i xifres. Registre de Càncer de Tarragona, Fundació Lliga per a la Investigació i Prevenció del Càncer.

[13] Galceran J, Ameijide A, Cardó X, Piñol JL, Gumà J, Saladie F, Izquierdo A, Marcos R, Moreno V, Borràs JM, Bosch FX, Viladiu P, Borràs J. (2008). El càncer a Tarragona, 1980-2001. Incidència, mortalitat, supervivència i prevalença. Registre de Càncer de Tarragona. Fundació Lliga per a la Investigació i Prevenció del Càncer.

[14] Gerber, F., Furrer, R. (2015). Pitfalls in the Implementation of Bayesian Hierarchical Modeling of Areal Count Data: An Illustration Using BYM and Leroux Models. Journal of Statistical Software, 63, 1–32. https://doi.org/10.18637/jss.v063.c01.

[15] Jensen O.M., Parkin D.M., MacLennan R., Muir C.S., Skeet R.G. (1991) Cancer registration. Principles and methods. IARC Scientific Publication No. 95.

[16] Lawson, A. B. (2018). Bayesian Disease Mapping Third Edition.

[17] Lawson, A., Lee, D. (2017). Chapter 16: Bayesian Disease Mapping for Public Health. In Handbook of Statistics (1st ed., Vol. 36, pp. 443–481). Elsevier B.V. https://doi.org/10.1016/bs.host.2017.05.001.

[18] Lee, D., Rushworth, A., Napier, G. (2018). Spatio-temporal Areal Unit Modeling in R with Conditional Autoregressive Priors Using the CARBayesST Package. Journal of Statistical Software, 84(9). https://doi.org/10.18637/jss.v084.i09.

[19] Lee, D. (2013). CARBayes: An R package for Bayesian spatial modeling with conditional autoregressive priors. Journal of Statistical Software, 55(13), 1–24. https://doi.org/10.18637/jss.v055.i13.

[20] MacLennan R., Muir C.S., Steinitz R., Winkler A. (1978). Cancer registration and its techniques. IARC Scientific Publication No. 21.

[21] Mesa Páez, L. O., Rivera Lozano, M., Romero Davila, J. A. (2011). Descripción general de la Inferencia Bayesiana y sus aplicaciones en los procesos de gestión. La Simulación Al Servicio de La Academia, 2, 1–28.

[22] Moran, P. A. (1950). Notes on continuous stochastic phenomena. Biometrika Trust, 37(1–2), 17–23. https://doi.org/10.1093/biomet/37.1-2.17.

[23] Paradis, E., Blomberg, S., Bolker, B., Brown, J., Claramunt, S., Claude, J., et al. (2020). Package "ape".

[24] Parkin D.M., Chen V.W., Ferlay J., Galceran J., Storm H.H., Whelan SL. (1994) Comparability and quality control in Cancer registration. IARC Technical report No. 19.

[25] Registre Mortalitat de Catalunya. Anàlisi de la mortalitat a Catalunya 2017. (2019). Departament de Salut, Generalitat de Catalunya.

[26] Rosich Solé, C. (2018). Evolució dels algorismes MCMC i la seva implementació en programació probabilística. Trabajo Fin de Grado. Facultad de Matemáticas e Informática, UB.

[27] Schoenbach, V. J. (1999). 6. Estandarización de tasas y razones. Comprendiendo Los Fundamentos de La Epidemiogía: Un Texto En Desarrollo, 129–152. www.sph.unc.edu/courses/EPID168.

[28] Vranckx, M., Neyens, T., Faes, C. (2019). Comparison of different software implementations for spatial disease mapping. Spatial and Spatio-Temporal Epidemiology, 31. https://doi.org/10.1016/j.sste.2019.100302.

8. Apéndice I

8.1. Conceptos básicos de teoría de probabilidad

8.1.1. Conceptos previos al Teorema de Bayes

El Teorema de Bayes, tambien conocido como el teorema de las causas está estrechamente relacionado con la probabilidad condicional. Así, a continuación se presentarán una serie de definiciones y resultados necesarios para comprender el Teorema y entender su forma.

Definición 8.1. ***σ -álgebra***
Sea σ un conjunto y $\mathcal{A} \subset \mathcal{P}(\Omega)$. $\mathcal{A}$ es una σ-álgebra de $\mathcal{P}(\Omega)$ si satisface:

- $\Omega \in \mathcal{A}$.
- $A \in \mathcal{A} \Rightarrow A^c \in \mathcal{A}$.
- $\{A_n\}_{n\geq 1} \subset \mathcal{A} \Rightarrow \bigcup_{n=1}^{\infty} A_i \in \mathcal{A}$.

Definición 8.2. ***Espacio medible***
Un espacio medible es un par, $(\Omega, \mathcal{A})$, formado por un conjunto arbitrario, Ω, y una clase de conjuntos $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ con estructura de σ-álgebra.

Definición 8.3. ***Espacio de probabilidad***
Un espacio de probabilidad es una terna $(\Omega, \mathcal{A}, P)$ donde:

1. *Ω es el espacio muestral: conjunto que contiene todos los resultados. El espacio de probabilidad será finito cuando Ω sea finito.*
2. *$\mathcal{A} \subset \mathcal{P}(\Omega)$ tiene la estructura de una σ-álgebra que nos permite describir todos los posibles sucesos.*
3. *La probabilidad P es una aplicación $P: \mathcal{A} \to [0,1]$ tal que:*
   - *$P(\Omega)=1$.*
   - *σ-additividad: $\{A_n\}_{n>0} \subset \mathcal{A}$ disjuntos dos a dos $\Rightarrow P(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n)$.*

Observación 8.4. Un espacio de probabilidad es un caso concreto de espacio medible.

Definición 8.5. ***Suceso o Evento***
Un suceso es un subconjunto del espacio muestral Ω asociado a un experimento ϵ.

Proposición 8.6. *Sean A y B sucesos en* Ω*,*

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Definición 8.7. ***Sucesos independientes***
Dos sucesos A y B son independientes si $P(A \cap B) = P(A)P(B)$*. Se notarán como* $P(A \cap B) = P(A,B)$*.*

Definición 8.8. ***Sucesos complementarios***
Dos sucesos A y B son complementarios si $A \cup B = \Omega$*. Sea B el suceso complementario de A, se nota* $B = A^c$*.*

Proposición 8.9. *Para todo suceso A en* Ω*,*

$$P(A) + P(A^c) = 1.$$

Definición 8.10. ***Probabilidad condicionada***
La probabilidad condicionada de $A \subset \mathcal{A}$ *condicionada por* $B \subset \mathcal{A}$ *con* $P(B) > 0$*, se define como:*

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{8.1}$$

Proposición 8.11. ***Fórmula de probabilidades compuestas***
Sean $A_1, \ldots, A_n \subset \mathcal{A}$ *con* $P(A_1 \cap \ldots \cap A_n)$*,*

$$P(\bigcap_{i=1}^{n} A_i) = P(A_1)P(A_2|A_1)\ldots P(A_n|\bigcap_{i=1}^{n-1} A_i). \tag{8.2}$$

Definición 8.12. ***Partición del espacio muestral***
Sean $A_1, \ldots, A_n$ *sucesos de* Ω*, se dice que* $\{A_1, \ldots, A_n\}$ *representa una partición del espacio muestral* Ω *si:*

1. $A_i \cap A_j = \emptyset$ *si* $i \neq j$*,*
2. $\bigcup_{i=1}^{n} A_i = \Omega$*,*
3. $P(A_i) > 0 \ \forall i$*.*

Proposición 8.13. ***Fórmula probabilidades totales]***
Sea $\{A_1, \ldots, A_n\}$ *una partición de* Ω *con* $P(A_i) > 0 \ \forall i$*. Sea* $B \in \mathcal{A}$*:*

$$P(B) = \sum_{i=1}^{n} P(B \cap A_i) = \sum_{i=1}^{n} P(B|A_i)P(A_i). \tag{8.3}$$

8.1.2. Teorema de Bayes

El Teorema de Bayes expresa la probabilidad condicionada de un suceso A dado un suceso B en términos de la distribución de la probabilidad del suceso B dado A y la distribución de la probabilidad marginal de A.

Teorema 8.14. ***Fórmula de Bayes***
Sean A, B $\subset \mathcal{A}$*, P(A), P(B)>0. Entonces:*

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{8.4}$$

Considerando dos particiones de Ω $\{A_1,\ldots,A_n\}$ *y* $\{B_1,\ldots,B_m\}$ *de conjuntos de* $\mathcal{A}$ *de probabilidad no nula. Entonces:*

$$P(A_i|Bj) = \frac{P(B_j|A_i)P(A_i)}{P(B_j)} = \frac{P(B_j|A_i)P(A_i)}{\sum_{k=1}^{n} P(B_j|A_k)P(A_k)}. \tag{8.5}$$

8.2. Variables aleatorias

Definición 8.15. ***σ-álgebra de Borel***
La σ-álgebra de Borel es la σ-álgebra generada por los conjuntos abiertos y cerrados de $\mathbb{R}$*. Se notará como* $\mathcal{B}(\mathbb{R})$*.*

Definición 8.16. ***Variable aleatoria***
Una variable aleatoria es una aplicación X: $\Omega \to \mathbb{R}$ *definida en el espacio de probabilidad* $(\Omega, \mathcal{A}, \mathcal{P})$*, tal que:*

$$X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{A}, \qquad \forall B \in \mathcal{B}(\mathbb{R}).$$

Definición 8.17. ***Ley de masa de probabilidad***
La ley de una variable aleatoria X es la probabilidad Q definida sobre $(R, \mathcal{B}(\mathbb{R}))$ *tal que* $\forall B \in \mathcal{B}(\mathbb{R})$*:*

$$\begin{array}{rccl} Q\colon & \mathcal{B}(\mathbb{R}) & \longrightarrow & [0,1], \\ & B & \mapsto & P(X^{-1}(B)) \end{array}$$

Q será la medida imagen de P por X :

$$Q = P \circ X^{-1} = P_X.$$

Definición 8.18. ***Función de distribución de una ley de probabilidad***
La función de distribución asociada a una variable aleatoria X se define como funcion f:

$$\begin{array}{rccl} F\colon & \mathbb{R} & \longrightarrow & [0,1]. \\ & x & \mapsto & P \circ X^{-1}((-\infty, x]) = P(X \leq x) \end{array}$$

Proposición 8.19. *Toda función de distribución F asociada a una variable aleatoria cumple las siguientes propiedades:*

- *es creciente,*
- *es continua por la derecha,*
- *satisface*

$$\lim_{x\to\infty} F(x) = 1$$
$$\lim_{x\to-\infty} F(x) = 0.$$

Definición 8.20. ***Variable aleatoria discreta***
Una variable aleatoria X será discreta si su ley está concentrada en un conjunto numerable $B_0 \in \mathcal{B}(\mathbb{R})$. Es decir, sea $B_0 = \{a_n\}_{n\geq 1}$, se define la variable aleatoria X como:

$$X = \sum_{n\geq 1} a_n \ \mathbb{1}_{A_n} \text{ donde } \{A_n\}_{n\geq 1} \subset \mathcal{A} \text{ partición.}$$

La ley de una variable aleatoria discreta se determina por los valores

$$P_X(\{a_n\}) = P(X = a_n) = (P \circ X^{-1})(a_n).$$

Definición 8.21. ***Funciónd de masa de probabilidad***
La función de masa es la probabilidad de cada valor o variable. Se define como:

$$\begin{array}{rccc} p\colon & B_0 & \longrightarrow & [0,1]. \\ & a_n & \mapsto & P_X(\{a_n\}) \end{array}$$

Definición 8.22. ***Función de densidad***
La función $f\colon \mathbb{R} \longrightarrow \mathbb{R}$ es una función de densidad si:

- *$f \geq 0$,*
- *f es integrable en el sentido de Riemann en $\mathbb{R}$,*
- *y satisface*

$$\int_{-\infty}^{\infty} f(x)dx = 1.$$

Definición 8.23. ***Variable aleatoria absolutamente continua***
Una variable aleatoria X es absolutamente continua su existe una función de densidad f tal que su función de distribución F se puede escribir como

$$F(x) = \int_{-\infty}^{x} f(y)dy, \qquad \forall x \in \mathbb{R}.$$

En este caso se dice que F es una función de distribución absolutamente contínua.